Thomas Schulze
Porto/ Berlin
2015

Advisor:
Ricardo Jorge Pinto

Universidade Fernando Pessoa

Data Journalism, Millennials & Social Networks

What does data journalism mean for journalists?
And how it can affect the Millennials?

Statement of Originality

I hereby certify that I, Thomas Schulze have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements.

This applies also to all graphics, tables and images included in the thesis.

The paper is presented to the University Fernando Pessoa as part of the requirements for degree of 2nd Cycle Master Degree in Journalism at the University Fernando Pessoa.

Signature:

ID No.: 29451

Place and Date: Berlin/ Porto, 16. September 2015

Abstract

Data is a term that is currently making massive waves amongst media and news media. With stories like the Iraq War Logs, it made its way onto the journalistic stage. Due to that, the purpose of this thesis is to investigate the characteristics of data journalism and its effects on journalism. Therefore, the related project is designed to make use of current models in order to find out what it takes through practical use. Furthermore, the conducted case study aims to determine the usage and acceptance by the audience of the social network Twitter and in particular the Millennials.

The subject of the study was the produced visual data outcome and the feedback given by citizens and especially consumers/ users of Twitter. The data for the study was gathered through a quantitative image type analysis and the record of retweets and favorites.

These data support the view that data journalism with its visual results appeal to the audience and the characteristics of the Millennials. Additionally, it was concluded that the basic attitude of journalists will not change but the tools and skills need to be implemented in the newsroom and the work process of the storytelling journalist that, as a result, will promote watchdog and citizen journalism.

Keywords: Data Journalism, Millennials, Social Networks, Watchdog Journalism, Citizen Journalism

Table of Content

List of Figures

List of Tables

1. Introduction

"*Journalists who master this will experience that building articles on facts and insights is a relief.*" (Gray, 2012, p.4)

1.1. Introduction

The task of journalism and media is more than keeping the citizens informed. They play a far more important role in the democratic societies by acting as a *watchdog* and monitoring the activities of public institutions and processes. Already in 1922, Lippmann discussed the role of media to form a public opinion stating that the news media represents the primary source for the public perception of the global events. In reference to this McCombs and Shaw (1972, p.176) supported the responsibility of mass media with their theory of the Agenda-Setting and further pointed out that "*mass media have a significant impact on our focus attention*" (McCombs, 1981, p.121). This responsibility of news media needs a certain amount of credibility, which it lost over the past. The Pew Research Center (2012) observed the development of the credibility of news organizations.

Although, the study puts its focus on the U.S. market, it can be said that this process also happens elsewhere and is a general phenomenon (Rosen, 2012), as for example in Germany. To this conclusion comes a recent survey made by *YouGov* for the online edition of the German newspaper *Die Zeit* (ZEIT ONLINE, 2014), which is about German media and the conflict with Russia and revealed that the trust in German news media is suffering as well. In detail, it points out that almost every second German (47%) mistrusts the media, believing that news media is biased towards politics. While only 40% think media is objective and independent and 13% do not have an opinion on the topic, or do not know whether they believe the news or not. Since this example is focused on a specific event, a survey by the think tank *GfK Verein* (2014) took a more general approach looking at the trust in professional groups. The report drew a similar picture and shows, only around four out of ten people (37%) trust journalists, and institutions that should have a higher reputation to be a *watchdog* over government, companies and other institutions.

Furthermore, with the appearance of the World Wide Web, the amount of information and data exploded and is rising every minute. For example, reading this first part takes around one minute. During this minute an amount of 1.354.440 GB (Pennystocks, 2015) was transferred over the Internet and the size is growing as you continue to read. To make this huge number more tangible the storage of one IPod touch with 128 GB would be filled with 18.286[1] songs in 0,00567 seconds. To put differently, during the six months that it took to write this thesis, more than 2.8 billion IPods were filled with over 51 trillion songs, which theoretically results in a playlists that could run for almost 322 million years without playing a song twice.

So, resulting the massive increase of information journalists have to adapt to make use of this big data, which has become increasingly important over the last two decades. A new tool to work with this digital information is *data journalism*. It represents a journalistic approach, which uses data to create stories and supports them with tables or visualizations. However, this new tool is challenging the traditional approaches to make and present news, because as Troy Thibodeaux (2015), editor at the Associated Press wrote "*data journalism describes neither a beat nor a particular medium* (…), *but rather an overlapping set of competencies drawn from disparate fields.*"

Before continuing with the thesis, the term (data) visualization in this paper is meant to describe all forms of data visualization, which also includes for example more complex infographics. To be more precise, all forms of visualization of data or visual elements to show a hierarchy or to differentiate. It is important to distinguish since paintings or photographs represent technically visualizations but are not meant and talked about.

1.1.1. Motivation and purpose of research

Conversations about data journalism easily show that the majority of people do not have a clear idea about what it is and usually guess the components it consists of - data.

Data journalism is considered a rather young discipline, which is tightly connected to a computational approach to journalism. Due to this young age, the data journalistic process is still described very vaguely, this field is in need for guidelines and support for journalist who are motivated to move and work in this field. However, data journalism

[1] Duration: 3:30, Bit rate: 256 kbps, Sample rate: 44,100 kHz

increased a lot in the past years and pioneers like *The New York Times* and *The Guardian* even created specialized websites and accounts on social networks to cover this fast evolving field. As a matter of fact, news media is trying to adopt and implement data journalism, which is currently very trendy, but represents a serious approach to regain the credibility and trustworthiness of journalism and inform the readers in an appealing way by making use of the characteristics of "*precision journalism*" (Meyer, 2002).

The importance of the Internet for data journalism was already mentioned but there is another element, which is challenging journalism - social networks such as Twitter or *Facebook*. In a study by Kwak, Lee, Park and Moon (2010) they conclude that especially news accounts on Twitter can be more powerful than traditional news media.

However, closely linked to social networks are in particular younger age groups like the Millennials and Generation Z. Due to that, a second focus is put on the relation between data journalism, social networks and Millennials. Hence, a research of the potential of the social network Twitter as a tool for data journalism and the feedback from the audience regarding data journalistic products was the point of interest.

1.1.2. Structure of the thesis

The structure consists of the three main parts: introduction, development and the conclusion that includes subchapters to simplify the reading experience and orientation. Furthermore to increase the reading experience each chapter is summarized in a short in the end of each chapter.

In the introduction, the reasons, approach, idea and motivation to do research in the field and write about data journalism and data visualization, are stated. In fact, the field of data journalism, as it is currently being used can be considered a young discipline, the main body or development covers different parts and presents the basic facts and knowledge that are necessary to understand the topic. The development, is divided into two main parts:

1. Data journalism
   - *Theoretical framework*
   - *Example of use - Dark Horse Vietnam*
2. Current state of data journalism on social networks
   - *Theoretical framework*
   - *Case study - data visualization on Twitter*

Part 1 is a represents the theoretical framework to data journalism by presenting, describing and explaining approaches and terms linked to data journalism and visualization. In accordance to these models the example of use, the data story *Dark Horse Vietnam* is put into practice and what this new form means for journalism.

Within *part 2* the focus is put on social networks and the characteristics of the so-called Millennials to provide the fundament for the case study. It considers the power of social networks and mobile application and investigates the use of data visualizations on Twitter by three pioneering newspapers in the field of data visualization.

The final stage of the thesis, the conclusion, picks up the main points and arguments to close the circle in order to answer the research questions and give an outlook. Moreover, perspectives of data journalism are discussed and brought into perspective according to findings of this thesis and other sources.

1.1.3. Research method

The general research method to this thesis was similar to a journalistic story, by using the strategy of concentric circles. That means relevant literature, important authors and key player in the field of data journalism were investigated and used to build a theoretical framework for this thesis (Krämer, 1999, pp.33-48). Additionally, in order to reach a conclusion for the thesis an inductive reasoning was used. The aim of the use was to explore whether data journalism and its visual products offer a way to reach and involve the audience, in particular Millennials. Due to that, the overall method used for this thesis is the concept of the grounded theory, which is all about data collection and analysis. The aim of this strategy is to construct a theory that is grounded in the data (Glaser & Strauss, 1967). The systematic process for qualitative research has an exploratory approach and its assumption that ideas emerge from data or information. Therefore, different studies were taken into account, in order to examine the development and answer the research questions. Meanwhile, the reasoning bases on three parts and observations made during the research. The first part is the theoretical part, in which studies and models are presented, while the second part puts the models to the test in the example of use *Dark Horse Vietnam*. Finally, the third part explores the consumer but also the producing side with the case study, which takes the theoretical framework into consideration. The case study is a social network analysis about the use of data visualization on Twitter by three different news media representatives. It furthermore is a quantitative image type analysis that focuses on the visual data journalistic products. The categorization into image types is a central component of this method (Grittmann, 2001, p.277) and the priority is put on the message of the image (Grittmann & Ammann, 2009, p.144).

1.1.4. Research question

As a matter of the fact, that this thesis combines two projects, the approaches are two different ones. The purpose of the first part is to gain an insight into the current state of data journalism, while making use of the research results in a practical form. The aim was to investigate required skills and knowledge in order to engage in data journalism.

Meanwhile, the focus of the second part is the attractiveness of images or more specifically data visualizations in combination with their use by three major news companies on the social network Twitter. The goal was to find out what the current state of the use is, while the question was raised to what extend news companies already make use of this possibility. In order to, find out if data visualizations can be used as a tool to spread information and knowledge on social networks and appeal to the audience, the main question was: How did the promotion and feedback of the visual results of data journalism develop on Twitter over the last years?

Finally, the thesis presents various reasons why data journalism should be adopted by journalists and news organizations. In addition, it will look at the role of Millennials in connection to data journalism and its development. Furthermore, an outlook on the perspectives and future of data journalism is given.

1.1.5. Research boundaries

During the research and gathering of data different boundaries, problems and difficulties appeared. The biggest obstacle that had to be overcome was to develop a reliable parameter to filter out data visualization. Since, the case study focuses on visual content the research happened mostly on a manual base. This procedure is prone to errors due to a human error. As a result of this problem, the gathered data was double-checked. The reason why no reliable API or search parameter could be developed was that the content of images could not be filtered. Furthermore, some data visualizations were included within the tweet as a Link.

2. Development

In the following, the principles and models of data journalism are presented and explained to form a fundament and understand what data journalism is about and why it matters (Part 1). Furthermore, the elements of the conducted case study are clarified; the study is further described and finally analyzed (Part 2). In addition, the work process with the different tasks was recorded and visualized and can be seen in *the Appendix 1*.

Part I

This part puts the focus on data journalism and provides the theoretical framework for a better understanding of the term and to put the data story *Dark Horse Vietnam* into practice.

2.1. Data Journalism

Data became an important element in our everyday life and is all-around, for example when taking the metro, shopping online, using loyalty cards, paying by credit card or simply visiting the doctor. In all of these cases, chances are high that, data is collected and analyzed by software, which is why big data became significantly important to understand customers, clients or patients. Moreover, the possession of such big data that describes large and complex sets of data, means potential revenue for different companies. In other words and to refer to the example of IPods, big data is for example a dataset of all 51 trillion songs with information about Artist, Track, Album, Genre and Duration of each song, which leads to a spreadsheet with more than 256[2] trillion columns.

Nowadays, big data is also used in other areas for instance to support police forces to predict when and where criminal activity is most likely to occur, but also intelligence and secret services such as the German B.N.D.[3] or the US American N.S.A.[4] make use of big data as the example of *WikiLeaks* shows. Up until now, its publications are of high public interest and show the significance of data these days. As a result of this development, an urge for a specialization in journalism that is able to deal and process such information with a journalistic sense appeared - data journalism.

Data journalism has already proved its right of existence for example when Julian Assange's *WikiLeaks* published the U.S. military logs of the war in Afghanistan (*Afghan War Diary* or the *Iraq War Logs*). Another example is the origin of the still ongoing N.S.A. surveillance scandal, which is part of the global surveillance disclosures leaked by Edward Snowden and first published by *The Washington Post* and *The Guardian*. In accordance to these revelations, *WikiLeaks* creator Julian Paul Assange picked up this new situation for journalism and stated in an interview with the U.S. American magazine *The New Yorker* that he wants "*to set up a new standard: scientific journalism*" (Khatchadourian, 2010). The term means, referring to Assange, that journalists should present and publish the data, which forms the fundament of the research and story. He further emphasizes the imbalance of power in which the audience is not able to verify the news or what they are being told, which leads to an abuse of the power by the journalists amongst others. So, by giving the background information, the audience can replicate, check and verify the data. This approach to journalism is strongly related to Philip Meyer's idea of precision journalism that is explained after the terminology of data journalism.

[2] 256.337.444.571.434 columns
[3] Bundesnachrichtendienst
[4] National Security Agency

2.1.1. Terminology

Although, data journalism is often described as a quickly evolving and growing area (Gray, 2012), it is not well defined yet. Due to this, many approaches to define data journalism and opinions about what it is and which components it includes can be found.

Certainly, the term data journalism seems rather simple to define and in a broad sense it can be described as journalism done with data. For example Paul Bradshaw, author of *Scraping For Journalist* and online journalist, said that data journalism is a matter of approaching data and "*(...) the compilation of data is what defines it as an act of data journalism*" (Bradshaw, 2011). Furthermore he states, "[d]*ata can be the source of data journalism, or it can be the tool with which the story is told - or it can be both"* (Gray, 2012, p.3). So, it either begins with a question, which then needs data to be answered or a dataset that needs to be investigated and questioned.

Also Steve Doig, the Knight Chair in Journalism of the Walter Cronkite School of Journalism of Arizona State University, thought about the concept of data journalism and described it as another way of gathering information, formulating that (Remington, 2012):

> "*It's the equivalent of interviewing sources and looking at documents, except with data journalism you are essentially interviewing the data to let it tell you its secrets (...).*"

The Citizens Campaign provided another, more general, definition by stating that (Skowronski, n.d.):

> "*Data Journalism is when we use public data and statistics to tell a story. With the development of modern technology and an increase to access to information in a digital format - it is now possible to extract more information from public data.*"

In reference to one of the pioneers of data journalism and creator of *The Guardian Datablog*, Simon Rogers (2013, pp.277-278), data journalism has a variety of styles "*from visualisations to long-form articles*", which means it can be presented in words or in a graphical way. He further emphasizes that the fundament, which all data stories have in common are numbers and statistics with the goal to create a story from that.

These approaches already show the difficulties of journalists and experts to clearly define data journalism. In fact, the same problem occurred when the associate professor Cindy Royal and lecturer Dale Blasingame (both from the School of Journalism and

Mass Communication at the Texas State University) asked several journalists to define the term at the *International Symposium on Online Journalism* in Austin, Texas. The same Q and A showed that it was easier for the participants to name the elements and process of data journalism, which are both important in order to understand data journalism. According to Prof. Dr. Ralf Spiller and Prof. Dr. Stefan Weinacht (2014, pp.411-433), the core features of data journalism are: "*(...) the collection, analysis and preparation of digitized information with the aim of a journalistic publication"*. As a matter of fact, Simon Rogers (2013, pp.308-309) describes and summarizes the important content and parts of data journalism, which in *Open data journalism: how it can work best*:

1. <u>Expose the data behind the story</u>
   Journalism should reveal something new about the world and be timely. By that he means that the audience and people care to become part of the story.
2. <u>Provide the key data people need</u>
   With all of the data available it is the journalists task to curate key numbers. Research and selection of data is an essential part a journalist's job and combining these skills helps the audience to find what they are looking for.
3. <u>Make it personal</u>
   The data should allow the user to see how it affects their lives. In particular, when the data is very "granular" the journalist can bring it to life by personalizing it.
4. <u>Anyone can do it</u>
   With all the free tools out there it is easy to visualize and analyze data. In the end, the key skill is to know whether something is a story or not.
5. <u>Make our data open</u>
   A journalist should publish the data in a format, which anyone can use.
6. <u>"Do what you do best, and link to the rest"</u>
   As Jeff Jarvis said "*There's bound to be someone out there doing something amazing - why not be open enough to embrace that?*" (Rogers, 2012, p.309)
7. <u>Free data now</u>
   Open data is important but real-time data became just as important for example when talking about election results.

8. We’re not the experts
   “*We can’t be experts in every aspect of life - why not try and engage those who are, so we can make them part of our process?*” (Rogers, 2013, p.309)
9. Make big data accessible
   It is the job of a data journalist to make the “big data” simpler, smaller and understandable.
10. Engage
    “*At the end of the day, it’s all about the stories*” (Ibid.)

These ten recommendations demonstrate the idea behind data journalism and the required elements to create a data story and work in the field of data journalism. Additionally, data journalism is often strongly connected to graphics and visualizations but also open data access. Although, the focus in this work is put on the visual approach of a data story, it is not essential to a data story. In the end, the most important part is about telling the story in the best way possible. That means, a news story can also be published with the spreadsheets of numbers because the choice of storytelling tools, depends on the story itself.

A huge benefit of using data, is in fact that it can enrich a story by giving insight and the possibility for the audience to explore and/ or even interact with the sources, so they can decide themselves which data they consider important and which not. Linked to this is the advantage to use data in order to analyze complex situations like political debates for example. The range of topics is huge, ranging from the financial sector over sports to social subjects. For instance, the production of an interactive, personalized calculator based on gathered data lets the citizens see how much of their taxes is spent on healthcare, for example. According to that, and by referring to data and numbers which is displayed in graphical ways, it helps to identify trends that otherwise might have stayed hidden.

Summary

Data journalism represents a special form of investigation that uses data to develop stories. Moreover, it is a specialized way to treat data as an information source and interrogate the researched material based on statistics. In the following, the findings are presented in a visual, sometimes graphical or in an interactive way to engage the audience. Furthermore, the term data journalism often goes along with open data journalism, which is based on *open data*. In addition, big and small data offers considering Callie

Neylan, Senior Designer at Microsoft, a unique insight "*(...) and the result of the tools that allow us to access, probe, poke, prod, dissect, visualize, and hopefully, make sense of it*" (Chiasson & Gregory, 2014, p.XII).

2.1.2. Precision journalism

Whenever reading or hearing about data journalism the term precision journalism is called out and referred to. The reason is the similar idea and approach to journalism, because the science sector merges powerful tools of data gathering and analysis. The gained data represents the most verifiable way to investigate and display the "truth" in a disciplined way. Hence, data journalism has an equal approach as precision journalism when gathering facts and creating stories because according to Meyer, it is (2002, p.235):

> "*(...) a way to expand the tool kit of the reporter to make topics that were previously inaccessible, or only crudely accessible, subject to journalistic scrutiny. It was especially useful in giving a hearing to minority and dissident groups that were struggling for representation.*"

Nevertheless, the idea of providing a good level of objective, scientific facts, in order to stay credible, is not a recent thought. In 1922 already, Walter Lippmann (1965, p.216) wrote "*The more points, then, at which any happening can be fixed, objectified, measured, named, the more points there are at which news can occur*". With this statement, he pointed out the importance and dependence of available objective facts and journalism. This approach to journalism and the idea of precision journalism is connected to Julian Assange's vision of a new standard - scientific journalism. He supports and actively empowers the idea of working scientifically and provides information or data, combining and putting them into context with journalistic storytelling skills.

Referring to Philip Meyer (2002, p.6), precision journalism was born in the 1970's, with the aim to move stories and journalism more towards scientific topics. The movement was motivated and supported by the increasing availability of computers, which could manage and process large amounts of data. Meyer also stated that this journalistic form was a response to Tom Wolfe's *new journalism*, which lacked objectivity and imposed personal viewpoints on the reader.

As a result, the scientific methods proposed by Meyer provide an objective and precise way to report about happenings and events. The data and information, which are gathered by making use of these social science methods, represent the fundament of data journalism because it includes practices for the data collection, documentation, analysis and precise presentation of the findings. Due to that the gathered datasets and numbers reach a higher credibility and accuracy and make a story less vulnerable and more transparent.

Summary

With precision journalism an ongoing trend has been established that gives reporters the chance to produce more fact oriented and less tenuous articles by using scientific methods. In order to produce stories, journalism can make use of tools from social science and scientific research methods, which are accepted as truthful. This is the reason why Meyer suggests, using and adopting scientific techniques and their "objectivity" to the process of mass communication, to increase depth and accuracy in journalistic stories.

2.1.3. Data journalism process

The process and components of data journalism were already mentioned and emphasized in the terminology. This part takes a practical viewpoint and describes different models and workflows to produce a data story. These were taken into account and used to create the article *Dark Horse Vietnam* (Appendix 7).

When engaging in data journalism, it is important to know how to correctly deal with data. The reason is that, considering Meyer (2002, p.6) it provides "*(...) the essence of precision journalism*", which results in quality data journalism. Meyer's idea of putting precision journalism into practice consists of two main phases: input and output phase. The goal during the input phase is to collect and analyze data, while the output phase is used to prepare the data for the "*(...) entry into the reader's mind*" (Ibid.). Furthermore, he divided the two main phases in six smaller steps, in which he states how to deal with data:

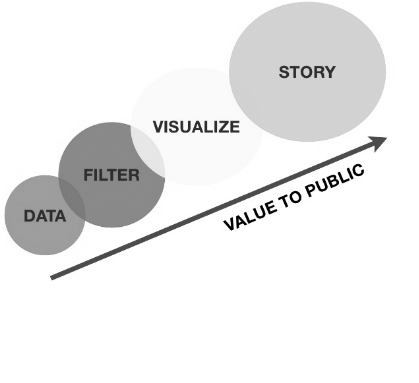
1. *Collect it*: You do not need to become a scientist but it is important to know their tricks and methods to collect data.
2. *Store it*: Choose a computer rather than stacks of paper to store your data.
3. *Retrieve it*: By making use of the tools of precision journalism, every kind of data can be retrieved, regardless the reason they were stored.

4. *Analyze it*: Find interesting anomalies, but look also for patterns, correlations and implied causations.
5. *Reduce it*: Reducing the data is as important as the data collection, to show what was included and what was left out.
6. *Communicate it*: "*A report unread or not understood is a report wasted.*" (Meyer, 2002, p.7)

Steve Doig's concept to tackle a data project is similar to Meyer's idea of the process of precision journalism. Doig gives three essential points, which need to be followed in order to understand and work with the data before a journalist can deliver the data or story. Referring to him (Gray, 2012, pp.153-156), the first step is to list questions and ideas that should be looked at before the data is requested and researched. After that, the collected data needs to be cleaned, since it still contains unnecessary data and will not answer the questions asked before. The last step, is that the fact that data may have undocumented features. He points out, that for example due to newly created codes some elements might not be documented in the data dictionary. In this case, the data dictionary, in this case represents an archive of all used components, their meaning and description.

Mirko Lorenz presented a similar approach to the process of creating a data story. In a presentation (Lorenz, 2010), which the information architect and journalist gave at the *7th Conference on Innovation Journalism*, he pointed out five main points in the data journalistic workflow process. Lorenz stated that the process starts with (1) digging into the (big) data and the need to gather and clean it, in order to give it a structure to work with. In the next step, (2) information should be mined and filtered to understand it and find patterns or anomalies. After accomplishing that, (3) the visualization process begins in which an adequate visual solution in a graphical or multimedia presentation has to be chosen. In the final two phases, (4) the product needs to connected to the classic storytelling to appeal to the audience and ultimately (5) media needs to be created, that provides a certain value for the audience. In addition, to this process, Lorenz (2010) also emphasized that each point has a certain value to the public, as it can be seen in fig. 1.

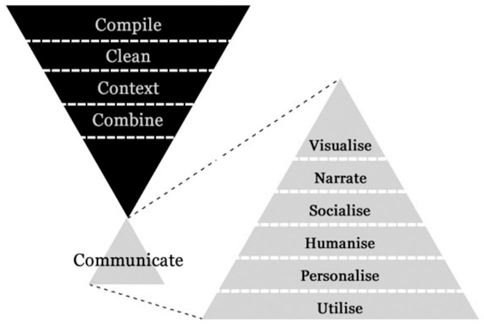
Fig. 1 - Data-Driven Journalism = Process Lorenz



The graphic summarized the process in four major phases, which have an increasing value to the public from *Data* to *Filter* over *Visualize* up to the *Story*.

Fig. 2 - The inverted pyramid of data journalism (complete)

Bradshaw



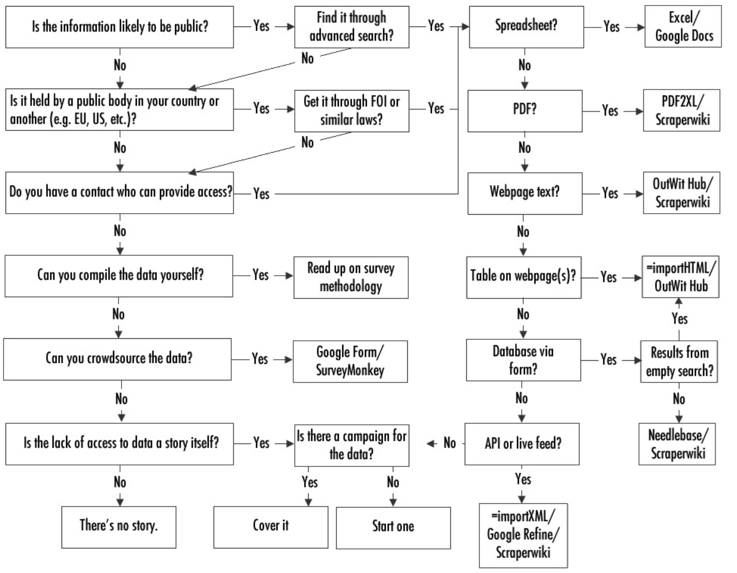
Another, more detailed model of the data journalism process was published by Paul Bradshaw, the *Inverted pyramid of data journalism* (fig. 2). With the model, Bradshaw supports the approaches described above by giving five important stages to make a compelling data story:

1. *Compile*: This is the fundamental and most important stage, since everything rests on the collection of information. Compiling data can take multiple forms and "*(...) defines it as an act of data journalism.*" (Bradshaw, 2011)
2. *Clean*: This step is needed to erase human errors and to convert the data into an interchangeable format to make it usable to everybody.
3. *Context*: It is important to have a clear question that helps to ensure not to lose focus and compile further data in order to contextualize them with the question.
4. *Combine*: Single datasets can offer a good story, but often a good story is the result of two or more datasets combined.
5. *Communicate*: It is important to visualize the results of the previous steps on for example charts, animations or infographics.

Step number five (*communicate*) leads to an extension of the model of the inverted pyramid, which Bradshaw describes as the pyramid of the communication process. It is still part of the model of the inverted pyramid and consist of six different types: (1) *Visualize*, (2) *Narrate*, (3) *Socialize*, (4) *Humanize*, (5) *Personalize* and (6) *Utilize*. Each of these types is explained more detailed in the chapter *visualization process*. In addition to the model, Bradshaw published a flowchart that puts its emphasis on "gathering data" (fig. 3). The graphic helps to identify the best ways of getting hold of data and dealing with it by asking the right questions.

Fig. 3 - Gathering data: a flow chart of tools and techniques

Bradshaw

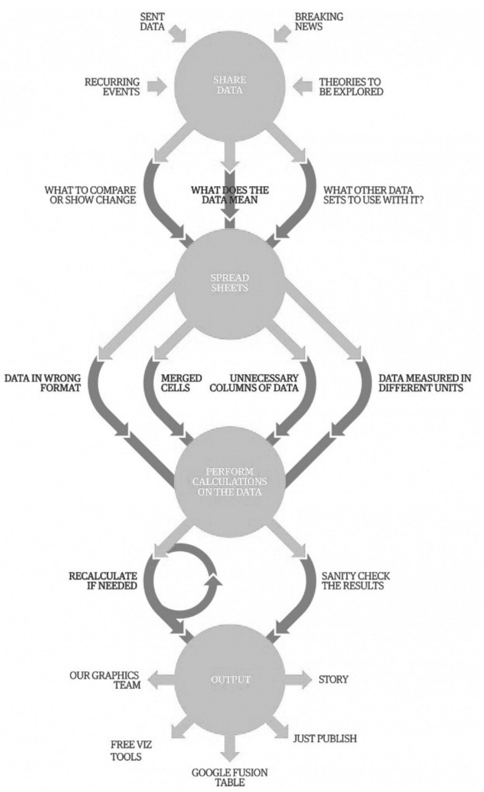

Is the information likely to be public?
Yes
Find it through advanced search?
Yes
Spreadsheet?
Yes
Excel/ Google Docs
No
No
No
Is it held by a public body in your country or another (e.g. EU, US, etc.)?
Yes
Get it through FOI or similar laws?
Yes
PDF?
Yes
PDF2XL/ Scraperwiki
No
No
No
Do you have a contact who can provide access?
Yes
Webpage text?
Yes
OutWit Hub/ Scraperwiki
No
No
Can you compile the data yourself?
Yes
Read up on survey methodology
Table on webpage(s)?
Yes
=importHTML/ OutWit Hub
No
No
Yes
Can you crowdsource the data?
Yes
Google Form/ SurveyMonkey
Database via form?
Yes
Results from empty search?
No
No
No
Is the lack of access to data a story itself?
Yes
Is there a campaign for the data?
No
API or live feed?
Needlebase/ Scraperwiki
No
Yes
No
Yes
There's no story.
Cover it
Start one
=importXML/ Google Refine/ Scraperwiki

Also Simon Rogers (2013, p.288-289) published a model of the process to create a data story, which is based on his practical experience at *The Guardian*. This workflow shows the steps that happen in their newsroom and is displayed in fig. 4. The flowchart by Mark McCormick and Simon Rogers describes and displays the cycles of creating a data journalistic story at *The Guardian*. Simon Rogers (2011), summarized the points seen in the model and wrote in addition to the processes in the visualized in the workflow:

- *We locate the data or receive it from a variety of sources, from breaking news stories, government data, journalists' research and so on*
- *We then start looking at what we can do with the data - do we need to mash it up with another dataset? How can we show changes over time?*
- *Those spreadsheets often have to be seriously tidied up - all those extraneous columns and weirdly merged cells really don't help. And that's assuming it's not a PDF, the worst format for data known to humankind*
- *Now we're getting there. Next up we can actually start to perform the calculations that will tell us if there's a story or not - and then sanity check them to see if it just sounds wrong*
- *At the end of that process is the output - will it be a story or a graphic or a visualization, and what tools will we use?*

Fig. 4 - Workflow - The Guardian

McCormick & Rogers



After presenting the different models of the general process and real life practice workflows, it can be said the approaches, ideas and general mindset towards the process of data journalism and precision journalism is very similar. In all models, reliable data is the base, which needs to be put into perspective and into practice to create an appealing data story.

Summary

Overall, the models have all a similar approach and follow similar steps to tackle data and to create data stories. All of them start with gathering data, which is cleaned and filtered so it can be put into context in order to create a data journalistic product to communicate the story in the best and most appropriate way. It is crucial to keep the story that needs to be told relevant by keeping it as a common thread throughout the whole process. Nevertheless, all of the presented models and workflows might miss specific tools and be incomplete. Due to this, the charts are constantly discussed, updated and should be considered as approaches, which help to create a data journalistic product.

2.1.4. Open data

The importance of open data to data journalism was mentioned briefly before when referring to Simon Rogers model of open data journalism. It is essential because the access to data, platforms or applications is needed to get and scrape information to create a compelling data journalistic story.

The term open data combines a *"(...) technological meaning and its philosophical meaning - with a focus on raw, unprocessed information that allows individuals to reach their own conclusions*" (Robinson and Yu, 2012, p.189). They also added that in the past scientist used this term in reference to unprocessed and raw data. The Open Knowledge Foundation (n.d.) agrees with the approach of *open data* and defines it like this: "*Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike*". The Handbook further summarizes the most important details in the following points and emphasizes the importance of the compatibility of different systems and organizations to interoperate with each other. As in the case of *open data*, it is the ability to work and interoperate with different datasets. In order to do so, they need to be:

- *available and accessible*: The whole data must be available. Preferably by downloading it in a convenient and modifiable form over the Internet.
- *re-useable and redistributable*: They should have permission to be re-used and re-distributed and also enriched by mixing it with other datasets.
- *universal participation*: Absolutely everyone must be able to access and work with the data. There should be no restrictions in non-commercial or commercial use.

In practice, this means there are plenty of stories out there in data formats, and just like Jonathan Stray of the *Overview Project* said: "*You don't need new data to make a scoop*". This statement already shows the process of democratization that is currently happening, and has happened over the last years. It further indicates the need to democratize data by making it available and easily accessible for the citizens or audience. Thereby, the audience can assure that the data, which forms the basis of the stories, is not biased in order to reach a specific conclusion. In fact, the publication of the used data is important to interact with the reader and get feedback or helpful tips. Subsequently, this input by the audience, might lead to follow up stories, which encourages readers and increases the chance that the readers return to consume more news products and/ or provide new sources. Especially as students or teachers sometimes have access to materials and can provide sources that are interesting to research and can add value or a new angle to a story. Therefore, it is important to make the data used for a story available and open. Currently, most of these sources are public and governmental institutions, which are obliged to publish data so the citizens can access them. One source for institutional data is for example *Eurostat* by the European Commission. Other than that, the past also showed that insider can provide data, which are then processed and published like the *SIGACTS* by *WikiLeaks*, for example.

Nevertheless, sometimes it happens that data cannot be downloaded from an open data source or the institution of interest. In these cases, there are different other possibilities to get the information and data, in case it was recorded. The easiest way is to simply ask for the information at the public institution, which could keep the data of interest. If the public or governmental institution recorded the wanted data, these datasets need to be accessible for everybody and have to be published. This also can be forced by making use of a Freedom of Information (FOI) request. A FOI of request can be separated in three main types. A request for (1) an entire database, (2) an unpublished dataset, which was collected by a public body and (3) to authorities to collect data on your own (Egawhary & O'Murchu, 2012).

Summary

In general, it can be said that the essential meaning of open is free and in connection to open data, this means free access to information and knowledge, which can be used, modified and shared amongst the users. Furthermore, it represents the foundation for open data journalism and provides the components for the story.

2.1.5. Problems of data journalism

Data and precision journalism can also cause problems if done wrong because all too often people and journalists just blindly believe data without questioning its quality. Paul Bradshaw also addresses this point in the phase *context*, by stating that information and data cannot always be trusted and further, "[l]*ike any source, it should be treated with skepticism; and like any tool, we should be conscious of how it can shape and restrict the stories that are created with it"* (Gray, 2012, p.3). Moreover, he states the information and data have a specific background, which is biased and follows an objective.

> "*So like any source, you need to ask questions of it: who gathered it, when, and for what purpose? How was it gathered? (The methodology). What exactly do they mean by that? You will also need to understand jargon, such as codes that represent categories, classifications or locations, and specialist terminology*."

Ben Goldacre (2008) presents the "cherry picking" in science and ways to support their ideas of a wanted result of a case or trial, which shows that data is not completely objective. In fact, it is important to look at and approach data with a certain amount of skepticism and look for its origin. Therefore, journalists need to check the data like they check their sources and understand scientific methodologies.

However, there are other problems that can appear when looking at verification of data sources. The main risks that Alejandro (2010, p.24) mentions for newsgathering and distribution concern *"(...) the accuracy, the need for verification and the loss of control over the information*". In particular, when working with social media this problem appears when scraping data from social networks (Silverman, 2014, p.8). Up until now, the public relied on news organizations, governmental agencies or other official sources to get credible information. But, in recent years these organizations increasingly turn towards the public as a source for information, opinions and perspectives. In order to, achieve a virtuous and credible communication circle, journalists need to triangulate different sources and verify conflicting information. The process, to verify information from blogs or tweets is often crucial, especially during natural disasters, for instance. Live reports posted during catastrophes are usually liable to misperception due to fear or panic, which might pervade all sides and/ or perspectives (Silverman, 2014). In general, social media leads to events but journalists have to track down the details the old-fashioned way by getting in touch with institutions, for example. Just like in traditional journalism, you need to ask the right questions to overcome most problems of data. This approach was also supported by a presentation of the data journalist Henk van Ess, when mentioned three main questions that should be asked. (1) What are the data rules are? (2) What and why is something missing? And finally, (3) did they change the rules?

These questions apply to the data provided by many institutions and companies. In van Ess's example (2013), he refers to a statistic and article (Erny, 2013), which shows the decrease in train accidents in Switzerland. But this change is a result of a change in definition of a train accident by the Swiss Ministry of Transport. To be more specific, accidents in Switzerland are not listed when the minimal damage is lower than a specific counter value. This amount increased from 1975 to 2008 from 5.000 to 100.000 CHF, which consequently resulted in a massive decrease of train accidents from 2007 to 2008. This picks up and supports Bradshaw's idea to question the information, in order to understand the data and put it into context.

Unfortunately, the verification of data collides with a problem stated by Steve Doig (Remington, 2012), which is that journalist all too often make mathematical mistakes, for example when mixing up correlation and causation. In this case, it is important to know or learn that correlation does not imply causation. He reasons these kinds of errors with the "*math-clumsiness*" (Remington, 2012) of journalists and recommends reporters and journalists to have a basic understanding of math, to find and tell a data story in the correct way.

Summary

It is important to understand data and approach it skeptically as the train accident example shows. Therefore, given facts or data should not be relied on blindly and data needs to be questioned. As with every source it needs to be verified and not all data automatically implies facts. Due to that, it is crucial to be able to deal with data and have a general understanding of scientific approaches. If this is achieved, the journalist is able to stay credible by using reliable data.

2.1.6. History of data journalism

Journalism has been going through a continuous change over the last decades, not only regarding types of media that are used but also in reporting and dealing with the audience. After the shift in media from newspaper to radio and to television, methods and purposes of journalism needed to be adjusted and reinvented every time, to make them suitable for each type. The role of the media has drastically changed in the recent years, which will be picked up in the following parts that deal with the transition to digital journalism and the resulting evolution of data journalism.

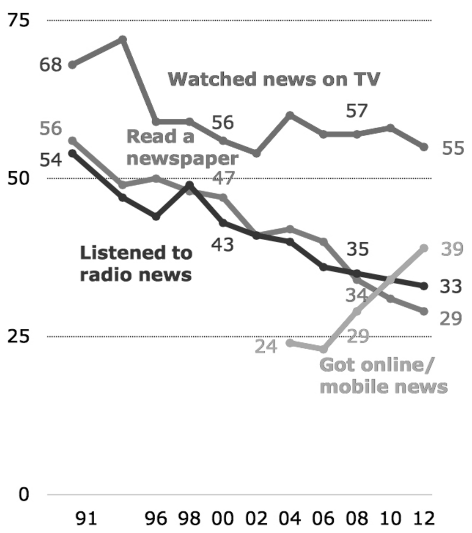
2.1.6.1. Transition to digital journalism

As mentioned, news media and press has changed throughout history and had to adjust to the new forms of media that also changed its role. While in the past, the press and in particular the printing press, saw its role as a gateway or gatekeeper that selects specific topics, which were then printed (Lippmann, 1965, p.192). This role has changed today, as a result of recent developments and the emergence of the Internet. Moreover, the web 2.0 and social networks also challenged traditional media as being a gatekeeper and sole source of objectivity, quality and authority (Beckett, 2008). As Jennifer Alejandro (2010) stated, the story about Michael Jackson's death of is an example of

> "*(...) how social media has breached the gap between traditional media and the consumers (audience). It proved that the gatekeeper role is no longer exclusive to journalists as the participatory culture of social media about engaging the audience has broken down the wall of journalism which separates the reader from the journalist*".

With the appearance of the Internet, another novelty occurred because the perception of news media by the audience changed significantly, from passive to active and currently a fragmented audience (Biocca, 1988). Furthermore, referring to Bardoel and Deuze (2001, pp.91-103) the audience became better educated as producer and consumer of content, which will lead to a citizen that is seeking information in a more direct and active way. The increased availability of computer and Internet supported the audience to become a producer and a consumer of news or content in general, which altered the landscape of journalism and minimized the relevance of the gatekeeping model. Henceforth, journalists were no longer the filtering element that had to pass on information that is worthy of discussion in the public sphere. As a matter of fact, the Internet enabled news way to interact, which changed the style of reporting but also opened up new resources for research. Moreover, it brought up new tools for journalists to communicate stories by producing images, videos or graphics and sharing them with a global community. The new opportunity to present a story supported by multimedia elements, brought a new narrative approach to tell a story in an appealing and engaging way. Furthermore, the digital technology affected and changed the news industry like the Pew Research Center (2012) report, *Trends in News Consumption: 1991-2012* shows (fig. 5). The survey shows, that news media like print, radio and even TV are becoming less attractive as a source for news for the audience. Also mobile news are rising and overtook print and radio news in 2012 as news source. Due to the decline in print editions and the circulation, some newspapers partnered their print edition with an online edition, while others focused fully on the online edition, eliminating its physical counterpart. Without getting into detail too much, this also affected the journalists since they had to adjust the style of writing to the online and/ or print edition. When presenting stories online, journalism no longer had to be linear and could break up the story into smaller parts. This gave the audience the possibility to explore and consume sections according to their interest. Subsequently, the web-centric approach became increasingly important, resulting particularly in the mobile consumption of news. This approach also resulted in a constant news flow because breaking news can be covered and send out in a 24/7 news cycle.

Fig. 5 - Where did you get news yesterday?



Accordingly, to the development of mobile news consumption, it is also unavoidable to get around ways to produce and consume news on mobile devices. This platform is getting increasingly important like a recent study on the use of smartphones in the U.S. (Smith, 2015) showed. In particular the Millennials are heavy users of portable devices. Wherever they go: 67% of this specific age group reportedly used smartphones to access the Internet. That compares with 40 percent of *baby boomers* and just 20 percent of *silents*. Furthermore, 82 percent said that they access the Internet with laptop computers and 47 percent with tablets (up from 26 percent in 2012), both significantly higher than older generations.

Nevertheless, the competition is huge when considering official websites of institutions or organizations and especially the attention social networks and media gets form the side of the consumers. In reference to the report mentioned before, 19% of all Americans have acquired news through a social networking site. Correspondingly, to the interactive nature of the web 2.0 and social media user started to produce and spread content on their blogs, forums or through social networking sites. Modern, digital journalism needed to adapt to this change by participating and working closer with the audience than before. The social networks were not only a place to publish or share news in order to pull user to the news site; it further became a source of news.

As a result of the immediacy of the social networks, journalism needed to adjust and accelerate the traditional journalistic process because the audience expects live and almost real time updates on events and happenings (Alejandro, 2010). Furthermore, the nature of the journalist has changed and the traditional roles such as photographer, editor or reporter are merging. So the digital journalist becomes a "*one-man band*" (Alejandro, 2010, p.34) that combines all of the roles in the production of stories for papers, websites, radio or TV.

The mentioned immediacy is further pushed forward due to the increase of news consumption on mobile devices (Smith, 2015). The reason is that mobile computing, the increase in connectivity and the powerful tools such as smartphones, tablets or phablets changed the way to gather, store and deliver stories. Due to that, it is unavoidable to get around ways to consume news on mobile devices and deliver live updates on events.

Summary

The major change resulting the emergence of the Internet is the changed role of the journalist but it also provided new tools and possibilities, which now need to be dealt with. The change towards digital media and journalism includes more elements, which cannot all be covered by this thesis and neither this chapter.

However, as stories and news become non-linear and new sources emerge, which are easy to access through the Internet, journalism has changed and will change. Moreover, news media organizations are no longer the monopoly of journalism and companies such as Google or Twitter will engage in competition with journalism. Due to that news companies try to adapt to this change and make use of the new opportunities of social media and the web 2.0 offer but they also raise the chances to engulf them in a professional crisis.

2.1.6.2. Evolution of data journalism

The Internet was a real game changer to the news media and has revolutionized the way information is published, stored and consumed. The emergence of the web and its growing importance also affected people's everyday life. Nowadays, almost everything is a collection of numbers - zeros and ones in binary codes. For example, the U.S. Postal Service (Nixon, 2013) produces big data by photographing and keeping record of every piece of paper mail from the outside. Another example, are electronic medical records (Robertson, 2013) in the healthcare sector that contain the complete medical history of a person or the mentioned mobile maps, which help people to navigate. In all of these cases, a big amount of raw data is produced, which is continuously growing and collected by different institutions. The development towards data journalism is closely related to the boost in information and technology over the past years, because the digitalization of almost every process in everyday life is recorded and this huge chunk of digital data and information offers new possibilities, which need to be explored.

The phenomenon of data journalism is not completely new and in the past, data journalism was not related to and based on computers and the Internet, as it is today. This proves the data story, published on the 5th of May 1821 by *The Guardian* (Manchester Guardian). The story is about the amount of students that received free education and the poor children that lived in the city. The story gets supported with a table that shows "*(...) a list of schools in Manchester and Salford, with how many pupils attended each one and average annual spending*" (Rogers, 2013, p.60, *Appendix 2*). So, neither data stories, nor the visualization of data are recent inventions like people such as Florence Nightingale show. She already made use of data and produced statistical graphics, which display the mortality of the British army in the east. The *history of data visualization* is dealt with in a chapter 2.2.2. and describes the historical development in detail.

Although data journalism is not completely new, it experienced a boost in its growth and importance due to the global interconnection by the Internet and the appearance of computers and their high performance in information processing. Moreover, it affected and changed the work and role of the journalist, because "[b]*y using data, the job of a journalist shifts its main focus from being the first ones to report to being the ones telling us what a certain development might actually mean*" (Gray, 2012, p.4). Furthermore, the interactive characteristic of the Internet offers the possibility for everybody to generate content, which consequently changed the involvement and relationship between journalist and the audience entirely. Before, the audience could only provide feedback writing letters or getting in touch via telephone. The new situation resulted in an increase of user-generated content from for example disasters like the *Typhoon Haiyan* or prior to that, the *Sumatra-Andaman* earthquake, which resulted in a massive tsunami in 2004.

In the following, the news and content were easier to access and distribute because social networks provided a network through which content could be shared and spread quickly. This leads back to the point of interaction between audience and journalist because both are "just a click away" and can communicate almost instantly.

With the increase of popularity of user generated content and the user-based or *citizen journalism*, the term and definition of a journalist also changed. The digital age and the Internet provide a possibility to gather and collect news and information without having to rely on individual journalists anymore, like the news site of *Google*. But, the characteristics of the Internet did not only change the consumption and perception of the audience, it also opened up a new resource for information and field to exploit. News were not limited to physical distribution anymore and could be published online. This novelty changed the way to communicate stories and put the journalist in a position to use different ways or channels to mediate news. Multimedia approaches in which text, videos, graphics and photos are used to tell a story like the recent example "*SEAL Team 6: A Secret History of Quiet Killings and Blurred Lines*" (Mazzetti et al., 2015) of *The New York Times* shows. In addition, interactive applications added a new way to communicate and personalize information and news.

A further change and advantage of the new virtual form to distribute made it possible to measure and analyze the consumption of articles on the online appearance of the news site. Furthermore, the social networks amongst other examples can be analyzed, like the case of the 2011 UK Riots (Rogers, 2013, pp.245-267) shows, to give background information on a current event. In addition, to these new possibilities and sources that the Internet offers are databases, which contain big data and huge amounts of information. This boost reinforced a more specialized form of journalism that was able to cope and work with this new resource.

Another aspect that empowered the development of data journalism was the idea of open data. In accordance to its progression data journalism experienced a push, when the *European Commission* initiated the so-called *PSI Directive - Directive 2003/98/EC* (European Commission, 2003, 2015). With the implementation, the member states started to provide data and information openly for re-use, which could be used by journalists to find stories and present them. By that journalists gained a new way to educate people, pass on knowledge and facts in form of datasets and visualizations. This open data approach is considered a fundamental component in (open) data journalism.

In consequence to, the change of journalism, the described style of journalistic work and the new resources, a revival of *watchdog journalism* was triggered. Because it is "*journalism that gives power to the people*" (Ward, 2014), referring to *Orlando Sentinel* editor Charlotte Hall. The *Poynter Institute* also refers to the conference *Creating A Watchdog Culture: Claiming An Essential Newspaper Role*, at which they stated that *watchdog journalism* is quality journalism, which needs to be better accessible and more exploitable. The Internet enabled an easier accessibility and higher connectivity, which also progressed data journalism further.

In the past days, data could only be published and accessed through books. Today, this data is provided in tables or spreadsheets, which can be accessed and processed by everybody with a computer and asking questions to the provided data. The increasing popularity of the Internet changed journalism in many ways but did not affect the work of journalists as drastically, considering a presentation by Simon Rogers (2015) on open data journalism. He points out that the reporters main tasks and skills have not changed at all by saying that the general work of a journalist is to, investigate, research, write and report, engage, reveal and expose. These tasks are also the main elements of the work of a data journalist. This applies to other tasks as well, which have not changed much for a data journalist either. The reason is, that for a good data storytelling, the traditional conventions of journalism are still needed and relevant because the journalist is a storyteller who needs to provide enough context, so the reader understands the story without forgetting to transport the primary message. In fact, this context, the verification of facts and checking of assumptions will continue to be a key component of the journalist's tasks. To bring it to a point, the primary task of journalists is considering Stephens (2014, p.XIII) to collect, present, interpret, or comment "*(...) upon the news for some portion of the public.*"

However, Spiller and Weinacht (2014) pointed out three major differences in comparison to traditional journalism, which are the significance of visualization, journalistic selection on a lower level and transparency of investigative results. In other words, the research process happens in a more technical, computational way when gathering data. Furthermore, the gained information or data is disclosed and displayed as data visualization, which is not the case in traditional journalism.

As a result of publishing the findings and raw data sets, the selection and interpretation of data is often left to the recipient when using interactive web applications like maps. Hence, the data journalist plays a less significant role as a gatekeeper compared to a traditional journalist.

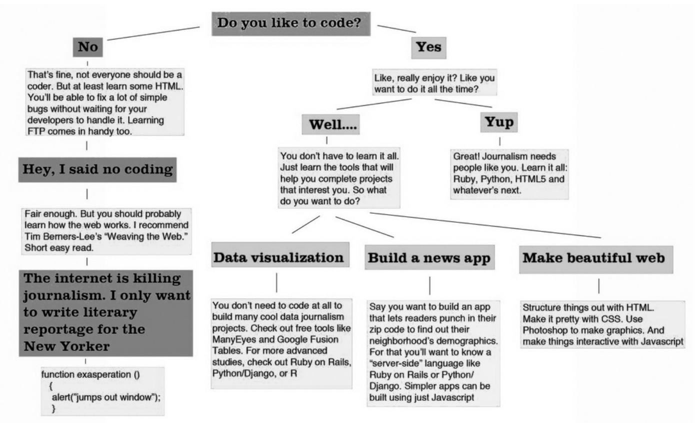
Another difference concerning required skills is a good sense for numbers, knowledge of social science and statistics. This sense goes beyond the basic 101 of math and aims at the skills related to programming and coding. Correspondingly to that, the question whether data journalist should be able to code or not, is often raised. Patrick Garvin, designer and graphic artist for the *Boston Globe*, as well as Simon Rogers (2015) and David Holmes (2013) approached the topic. In the article: *Should journalists learn to code?,* Holmes discusses the new situation and required skills, which Patrick Garvin picked up again in March 2015. Prior to giving the outcome of the articles, it is important to point out that in the past and still now, the process of producing data stories can be rather technical and complex in computational terms. In fact, skills tightly linked to coding are helpful because they help to transform and clean datasets, as well as provide the fundament to produce interactive visualizations or applications. Due to that, the discussion about whether journalists need coding skills or not, is important to consider.

The essence of all articles is that journalists do not need to be in full control and able to code but should have interest and a glimpse of knowledge of how programmer work in order to understand this tool and articulate their goals. In detail, David Holmes chose to provide a visual answer to the question - the *Coder flow chart*, which can be seen in fig. 6. Holmes agrees with and quotes the USC professor Robert Hernandez (2013) who stated,

> "*(…) not all journalists need to learn and master coding in JavaScript, Python, or Ruby. But they should know that it is not magic and, to be successful in their modern careers, they need to be able to communicate and work alongside different experts, including programmers. They need to be, at a minimum, digitally literate.*"

In the following, Simon Rogers picked up the flow chart, thinking about what it needs to teach data journalism. Referring to him, being a journalist is not about learning randomly technical skills, unless they are really needed in the newsroom. Additionally, he emphasizes that in 2015 it is significant to a journalist to be comfortable with data. For Rogers, the essential component for data journalism is to know how to tell a data-driven story, because the tools to find stories in data might change over time but the principles remain constant. Also, the Patrick Garvin considered this question and stated, that journalist should make use of technology to make their job easier and be better at it. Following this, he mentions it should be asked in how far coding or markup languages help to do a better job and make it easier to work for the journalist. With the first question Garvin picks up Rogers approach, supporting the idea that the technology should be corresponding to what is needed in the newsroom. Accordingly, it does not necessarily mean to have knowledge in coding, since technology that help to make the job of a journalist easier or better include software such as *Excel.* Garvin 's follow up question aims at the communication process between reporter and developer, because it is necessary when working on a project, both experts should have the same mindset and be able to communicate what they need from each other. Only in this case they are able to create a good data story using big data and interactive visualizations as a result.

Fig. 6 - Should journalists learn to code? Holmes


Do you like to code?
No
That's fine, not everyone should be a coder. But at least learn some HTML. You'll be able to fix a lot of simple bugs without waiting for your developers to handle it. Learning FTP comes in handy too.
Hey, I said no coding
Fair enough. But you should probably learn how the web works. I recommend Tim Berners-Lee's "Weaving the Web." Short easy read.
The internet is killing journalism. I only want to write literary reportage for the New Yorker
function exasperation () { alert("jumps out window"); }
Yes
Like, really enjoy it? Like you want to do it all the time?
Well....
You don't have to learn it all. Just learn the tools that will help you complete projects that interest you. So what do you want to do?
Yup
Great! Journalism needs people like you. Learn it all: Ruby, Python, HTML5 and whatever's next.
Data visualization
You don't need to code at all to build many cool data journalism projects. Check out free tools like ManyEyes and Google Fusion Tables. For more advanced studies, check out Ruby on Rails, Python/Django, or R
Build a news app
Say you want to build an app that lets readers punch in their zip code to find out their neighborhood's demographics. For that you'll want to know a "server-side" language like Ruby on Rails or Python/Django. Simpler apps can be built using just Javascript
Make beautiful web
Structure things out with HTML. Make it pretty with CSS. Use Photoshop to make graphics. And make things interactive with Javascript

To put it in a nutshell and refer to Rogers (2013, p.22), one can become a top coder but it is the bigger task to see what is new, interesting or correlates with other numbers or something else. Approaching numbers from that perspective is by far more important than the analysis, which Rogers proves regularly at *The Guardian* and the related *Datablog*. The platform engages the audience to participate like in 2012, when they asked for help to visualize the world's economic recovery. All the required data was provided, openly accessible and could be downloaded from the website. Until now, many more stories were published on the website and this approach appeals to the reader considering Rogers because "*(...) the average amount of time spent on a Datablog article is 6 minutes, compared to an average of 1 minute for the rest*" (Gray, 2012, p.139).

Summary

With the information and data explosion, as a result of the digitalization, the audience or citizen will be overrun with information and will not be able to filter and understand all of them. Due to that, "[g]*athering, filtering, and visualizing, what is happening beyond what the eye can see has a growing value*" (Gray, 2012, p.3) and needs to be picked up and done by journalists. The general mindset of a data journalist is still the same as the one of a traditional journalist, everything that changed are the tools used. Nevertheless, crucial skills for data journalists are, a sense for numbers, curiosity and most important the visual sense that is a significant part of the story and plays a major role in spreading the information and data. Although, the technology has changed, the urge to for compelling stories remains essential.

2.2. Visual communication - data visualization

There are many ways to visualize data, reaching from simple data visualization like a pie chart over more complex infographics to an interactive web application and as Stephen Few (2012, Preface) wrote: "*Good communication doesn't just happen; it is the result of good design*". A list of the most commonly used types of visualizations can be seen in *Appendix 6*.

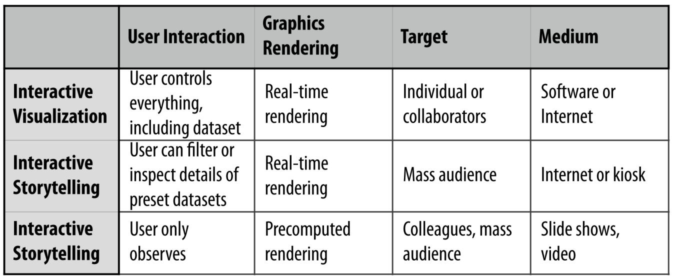
2.2.1. Terminology

In general, the term describes the presentation of data in a pictorial or graphical format (SAS Institute Inc., 2015) and is a product of data journalism that helps to engage the audience and *"(...) entices them to dig deeper into the content*" (Lankow, 2012). John C. Hart (2015) further emphasizes the relation to computer, stating: "*Data visualizations is a high bandwith connection between data on a computer system and a human brain, facilitated by visual communication*". Since, the connections and relations between data can be very complex and data visualizations offer an opportunity to portray these relationships in a unique way by illustrating the accurate "numbers". Moreover, the results enable the viewer to make out trends or detect patterns and henceforth derive and create a deeper insight (Chiasson & Gregory, 2014). In reference to Edward Tufte (2001, p.10), the good display of data represents processes and mechanisms, which help to make decisions and reach conclusions. That means, "[t]*he purpose of visualization is insight, not pictures*" (Card, Mackinlay & Shneiderman, 1999, p.6).

According to Hart (2015), there are three modes of visualization, *interactive visualization, presentation visualization* and *interactive storytelling*. The main purpose of the *interactive visualization* is to discover, which is used by a single investigator to plot the data. Whereas, the *presentation visualization* aims for communication and is produced to reach a large group or mass audience, in order to communicate some aspect of data. The main difference between the three modes (fig. 7) is the user input. In particular, the *presentation visualization* does not support any user input and the user can only observe the data. Furthermore, the Internet enabled a third mode of visualization (*interactive storytelling*), which can be found in between the two mentioned modes. This form is a presentation through interactive webpages that let user interact with data and investigate further.

Fig. 7 - Modes of visualization

Hart



	User Interaction	Graphics Rendering	Target	Medium
Interactive Visualization	User controls everything, including dataset	Real-time rendering	Individual or collaborators	Software or Internet
Interactive Storytelling	User can filter or inspect details of preset datasets	Real-time rendering	Mass audience	Internet or kiosk
Interactive Storytelling	User only observes	Precomputed rendering	Colleagues, mass audience	Slide shows, video

Summary

As more and more data is produced, collected and analyzed, the more important becomes the visual presentation of data, in order to understand, grasp the meaning of the data points and consequently communicate them (O'Reilly Radar Team, 2012). This communicative task of data visualization represents, to refer to Maria Popova (Chiasson & Gregory, 2014, p.VIII):

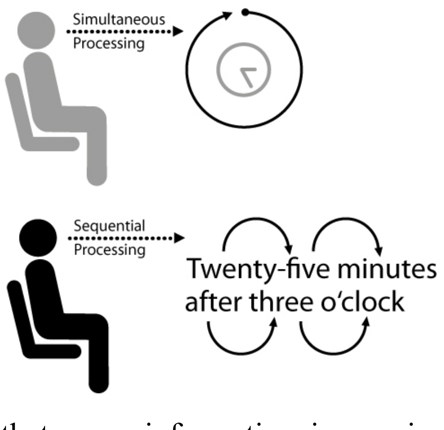
> “*(...) the intersection of art and algorithm, data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe.*”

2.2.2. History of data visualization

As Lankow and Ritchie wrote (2012, p.30): “*People have long accepted the notion that a picture can replace a thousand words, and similarly, that a simple graph can replace a table full of numbers*”. This statement shows the power of images because they are accepted as a credible source of information amongst the people. The reason for that is that the humans are a visual and also a symbolic species (Deacon, 1998, pp.215-217). This is a result of the process of cognition that builds on psychology, which is built on neurophysiology.

However, this chapter is not going to look closer at the interaction of subatomic particles and focus on the interaction between individuals instead. Referring to, Hutchins (Holyoak, 2012, p.48) the occurrence of human thinking in particular context is the starting point of situated cognition. This “thinking” originates in the human interaction by using cognitive tools and acting within social networks. Visualization became an important part of the cognitive system and is also increasing its meaning. In fact, “*the visual display provides the highest bandwidth channel from the computer to the human*” (Ware, 2004, p.2). According to, a research published by Professor Mriganka Sur (1996) of the MIT, “*(...) half of the brain is devoted directly or indirectly to vision, (...)*”, which supports Ware’s approach that more information is acquired through vision than all other senses combined. Smiciklas (2012, p.10) supported this argument by giving a visual example, which is presented in fig. 8.

Fig. 8 - Image versus Text
Smiciklas



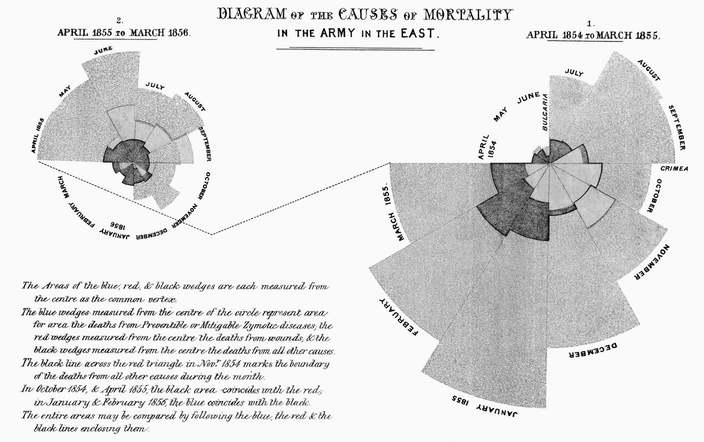
Nonetheless, it is also important to point out the change of the term visualization. According to Colin Ware, who referred to the *Shorter Oxford English dictionary* from

1972, visualization described until recently the process of creating a visual image in the mind. But, by now it became increasingly the representation of data and concepts that supports the decision making process.

The origin of visualizations, images and design of information has a long history that dates back to cave paintings. Since writing was not utilized back then, pictures were the best way to exchange and communicate ideas, which cave paintings all around the globe prove. Many thousand years later, around 3.000 BC the Egyptian hieroglyphics represent an example of using iconic and graphical language, which marks an early stage of information visualization. Around 4.500 years later, in 1350 the French philosopher Nicole d'Oresme created one of the first graphs. The data visualization of velocity shows and explains how moving objects with constant acceleration can be measured. The advantages of visualizing complex processes or structures were also considered by Leonardo da Vinci in the 15th century, for example when creating and illustrated guide on the anatomy of human. Later, during the Victorian Era in 1858, the efficiency of these images was proven again, with Florence Nightingale's infographic "*diagram of the causes of mortality in the army in the east*" (fig. 9). In her report about the *mortality of the british army*, she used statistical graphics to present data to the Parliament.

Fig. 9 - Diagram of the causes of mortality

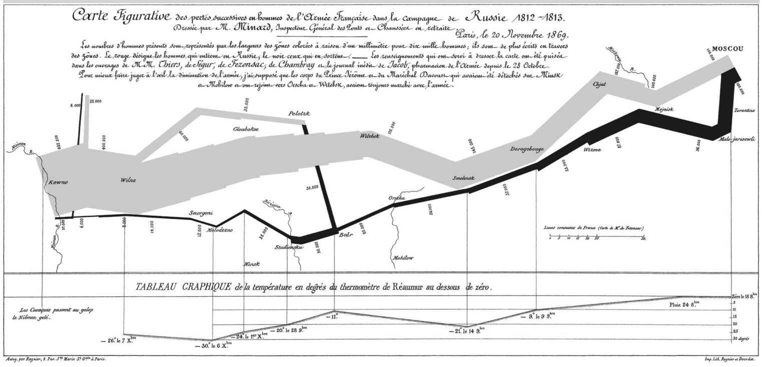
Nightingale



Even though Nicole d'Oresme already used bar charts, William Playfair is considered the person who introduced pie and bar charts in the book *The Commercial and Political Atlas* in 1786, to illustrate the Scottish trade from 1780 to 1781. Another meaningful visualization is, Charles Joseph Minard's graphic of the French invasion of Russia from 1812 to 1813 (fig. 10). It shows the casualties of the invasion by merging a timeline and charts. Bigger images of the visualizations can also be found in the *Appendix 3*, *4*.

Fig. 10 - French invasion of Russia

Minard



All of the examples represent visualizations that combine data and information about casualties in a graphic with a clear message. Prior to these exceptional examples, data and visualizations were mostly used to display the geographical world and make it understandable through maps. After that, the importance of infographic evolved slowly until 1930 to 40, when Otto Neurath developed a visual communication model called *Isotype* (Hartmann & Bauer, 2006). This model introduced an international picture language to communicate ideas and concepts through icons, pictographs and pictures, which simplified communication significantly and became a component of information graphics. From the 1970's onwards infographics became more popular until, according to Guy Kawasaki, the next important step happened in 1982 when *USA Today* decided to move away from "*(...) the text centric, black-and-white newspaper format and used color pictures and infographics to report news*" (Smiciklas, 2012, p.XIII). Kawasaki further mentions that others criticized this move, stating it would never last because it makes people stupid. Obviously, they were wrong because people and user like to see pictures and visualizations of information online on the web or in the printed newspapers or their e-reader.

Summary

In summary, visual communication has a long tradition like the 32.000-year-old cave paintings prove. Over time, data visualization and infographics progressed and proved and still do that up until now images rule as a way to communicate information.

2.2.3. Visualization of data

The visualization process in the data journalism process is used to analyze the data but also to communicate the results and answer the questions asked before. Since the second part of the thesis puts its focus on the visual aspect of the results of data journalism, the following presents the basic ideas of data visualizations. Additionally, a test whether a visualization is efficient or not is described. Data visualizations are omnipresent, no matter whether a pie or column chart on the daily news or an infographic in the online edition of a newspaper. But there are many errors that can be made as the chapter *problems of visualization* shows and not all of them are of high quality. Due to that, there are certain points to follow to ensure that the visualization is self-sufficient. Therefore, it should be clarified that this chapter is not a graphical guideline but for an explanation for general understanding.

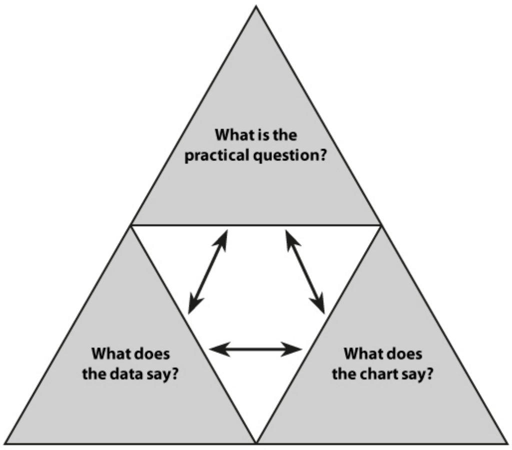
In general, it can be said that the wheel of visual storytelling does not need to be reinvented, because the history of print visualization remains relevant when creating online data visualizations. The ability to communicate in an incredibly effective way and provide a unique way of distributing the content is one of the major advantages of visualization. The primary rule when using visualizations is that it needs to follow the journalistic context and not the idea of a beautifully designed visualization, which does not represent the main idea. In addition, the visualization should be accurate, reasonable and balanced just like a non-data journalistic article. That means, according to the Chiasson and Gregory (2014, pp.147-158) asking the right questions before thinking about how the visualization should look. It is important to answer the questions that are expected to be asked by the reader. Therefore, determine the message to create a well-organized and compelling visualization.

Overall, the whole visualization should be unique but still keep the balance by making the visualization relevant and suitable with the topic and audience. This means understanding and knowing the audience, is a significant factor and the illustration should not overwhelm the readers and cover the key aspects. In order to test if the visualization is communicating the data efficiently, the data advisor Kaiser Fung introduced the idea of the *Trifecta Checkup* (fig. 11), which he presented at a talk at the NYU Stern School (Fung, 2010). The model represents a framework for chart critique and ensures that visualizations are effective, because as Fung (2010) stated, "[a]*ll outstanding charts have all three elements in harmony*". Another article (Fung, 2009) on his blog *junkcharts* gave a practical example, concerning the publish-worthiness of charts in connection to a chart made by Andrew Gelman (2009) that was transformed to a publishable version for *The New York Times*. The chart made by Gelman and the op-ed[5] creasted by the graphical department of *The New York Times*, is described more in

[5] opposite editorial page

Fig. 11 - Trifecta Checkup

Fung



detail in the *Appendix 5*, which also includes Fung's compiled list of changes that were made, in order to make a publish-worthy chart for the newspaper.

In relation to the *Trifecta Checkup*, Fung also named the, in his opinion, absolute worst practices in information visualization, which are too many charts that do not address a specific question. Furthermore, he points out the problem of using bad data and criticizes the use of "*pretty things that distort our perception*" (Bullock, 2012) by giving the example of bubble charts.

Summary

In summary of this part, it can be said that the information visualization is not a simple process and the creator should think of his target group and follow the principle of KISS - keep it simple and stupid or the principle of form follows function or to refer to Tuft (2006, p.51) "*simpleness of data and design = clarity of reading*".

Furthermore, the unique method to distribute the content in an instant way is a major strength but also a main weakness because it means that the audience might spend less time looking at it. As a matter of fact, data visualization is very effective for the distribution of information but not for the involvement of the audience. Due to that, data visualizations should contain a link to its source, which offers more to the people when they look it up and create an interaction between data and reader.

2.2.4. Visualization process

Since, the plain data is fairly unappealing in a spreadsheets, these numbers need to be communicated in a more attractive and visual way by using data visualizations. In the following, ways and models to communicate data stories are presented, which were also considered for the example of use.

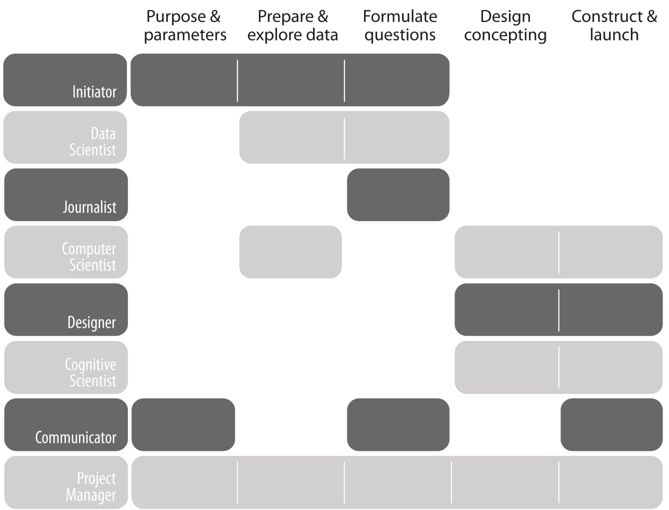
2.2.4.1. 8 hats of data visualization

The author of the book *Data Visualization: a successful design process*, Andy Kirk (2012) defines data visualization as "*(...) representation and presentation of data that exploits our visual perception abilities in order to amplify cognition*". Kirk went a step further in 2012, to look closer at the workflow and specialists involved in the process of data visualization and created the model of *8 Hats of Data Visualization* (fig. 12). The model puts its focus on the display of data. Nevertheless, it includes the collection and processing of data but narrows it down to the task of one specialist (data scientist). Due to that, this model can be considered a model for visualization. It describes the roles or mindsets and five cycles, from finding a story to putting it into practice. The result of this process is a story that is supported by the visualization of data.

The project starts with a *purpose & parameters* in which analytical directions of the project have to be established, the purpose, motive and parameters need to be determined. In the following two steps (*prepare & explore data* and *formulate questions*) the information and data is cleaned and mined to subsequently formulate questions, which need to be answered and communicated. In the final two phases, a design concept that is most suitable for the story is brought up (*design concepting*) and ultimately put into practice (*construct & launch*) by producing and launching for instance an interactive application. While the aim and tasks of the five cycles were described, the roles of the specialists still need to be clarified and are further explained to understand their specific tasks and field of expertise that coincide with the process.

Fig. 12 - 8 Hats of Data Visualization Design

Kirk



Initiator - "leader" - seeks a solution

Is the person who discovers a problem, is curious about a subject or just sees an opportunity? Either way, this person wants to explore a specific field and find answers. So, the mindset of the initiator is similar to a researcher that seeks evidence, defines the analytical emphasis and direction of the project. Furthermore, the person identifies and sets the framework or parameters to work with.

Data Scientist - "data miner" - acquires the data

The data scientist takes care of the data and its quality so it fits its purpose. For that the person prepares the data, enhances and consolidates it. Data miners need to have a very good amount of statistical knowledge and conduct initial descriptive analysis and exploratory visual analysis.

Journalist - "storyteller" - establishes narrative

The main task of the journalist is to formulate questions and look for stories and key positions and connect them with each other. In order to, find a good narrative the person needs to have the mindset of a deeper researcher that wants to get answers and validate the enquiry.

Computer Scientist - "executor" - brings the project alive

The capability to deal in a critical technical way, which means to acquire, handle and analyze the data, is the main task of the computer scientist. Moreover, the position requires technical illustration skills and technical programming skills to execute and "embody" the data gathered before.

Designer - "creative" - conceives the solution

The creative position needs to understand the message and possibilities. These then need to be explored and different options need to be pursued from a rational and reasonable design point of view.

To sum it up the balance of form and function is essential for the work of a designer.

Cognitive Scientist - "thinker" - visual perception

Knowledge of how eye and brain correspondent and work is the approach of the cognitive scientist. Also the principles of color theories, HCI-memory, attention and decision-making are important to this position.

Communicator - "negotiator" - needs a hard hat

The product needs to be communicated to the client, which is one of the tasks of the communicator. Other tasks are the gateway design, manage the expectations of the client and present the possibilities. In the final phase the launch, promotion and public communication of the product are tasks of the "negotiator".

Project Manager - "manager" - looks after the project

The person that is involved in every step of the process is the project manager. Fundamental to this position is to mange the progress and connect the different positions and steps. In order to, do that the person needs to understand the capabilities and finish, checks and pay attention to details. The manager is also concerned with the visualization and ethics in statistics. Moreover, the person identifies and sets the parameters together with the initiator.

Andy Kirk added to his model that he is aware, that labels and descriptions of the mindset and duties he gave are discussable and can be argued about. However, in conclusion it can be said that regardless of whether the project is done individually or in collaboration, the model helps to divide the process by duties and responsibilities. He emphasizes that the overall idea is to help to identify strengths and weaknesses when working on a data visualization project. Especially, the weaknesses can then be easily made out and worked on, or others can help to fill this knowledge gap in order to get the best data visualization result.

2.2.4.2. Inverted pyramid - communication

This part presents the "*6 ways of communicating data journalism*" (Bradshaw, 2011) and completes the model of the *inverted pyramid*, which was described before. Bradshaw distinguishes the process of "*communicating data"* (Ibid.) in six types that all provide ways to communicate data journalism and tell a story involving numbers. The six types that should act as a primer, according to Bradshaw are:

Visualization

Visualizing data represents the quickest and easiest way to communicate the outcome of data journalism but it does not naturally mean effectiveness. Especially, "chartjunk" shows that visualization can be affected by *churnalism* or become "*spectacle without insigh*t" (Bradshaw, 2011). Its major advantage is that it can make communication very effective, which is also a weakness since "*(...) people often do not spend much time looking at it*" (Ibid.).

Narration

Less is more - This applies for visualization but also for the narration. Here it is important that the writer thinks about the meaningfulness and the objective in communicating the numbers. In other words, abstract numbers or amount can be impressive, but without a meaning since the reader cannot relate to them. Therefore, it is important to reduce the amount to an understandable, manageable quantity such as the amount per person. And finally focus on the essentials by editing and link it to the whole.

Social communication

Communication is already a social act, and the success of infographics and the related data on social networks. In particular, *The Guardian* has cultivated a helpful and critical community around its data blog. Moreover, referring to Rogers the users stay longer on a piece on the blog than on a "normal" article on *The Guardian* website. But, the users are not only consumers; they also provide data as in crowdsourcing initiatives to obtain data, which shows another dimension of social communication. This indicates the new opportunity to present, and make use of data journalism, in a socially connected way. Furthermore, the social data based on usage and information gathered on different social networks, is a way of gaining data and also of communicating the results of data journalism, by making use of social dynamics - sharing, collaborating, campaigning. In fact, this field offers a new aspect for online journalism and data journalism; it is emphasized in the second part of this thesis linked to the case study of the use of data visualizations on Twitter.

Humanize

Stories based on numbers always present a challenge since they are difficult to illustrate. Tackling this challenge became easier, due to the emergence and growth of computer generated (motion) graphics (BBC Academy, 2015) that help to present abstract numbers and tell the story. This is the most important point, humanize the numbers and make them understandable to the audience. Simple examples are numbers, which extend the scales of everyday life and are beyond the level that "humans" can deal with. To engage the audience it is important to put the data into a representative case study and tell the story, for example, of an affected person besides giving the abstract data to the reader.

Personalize

In relation to, the point social communication another possibility opens up as a result of the changes towards online journalism - interactivity. For data journalism this means that the audience can choose and control which information they would like get presented based on their input. One example for interactive engagement, are forms in which the user can find out about how they are affected by a specific issue.

Other examples are geographical ways to get further information about a specific area or a combination of the approaches and personalization of a story by making use of the data provided by the user on social networks or other third party tools. This again demonstrates the important role of social networks and points out ways in which social strategies and personalization can be combined and the perspectives they offer for data journalism.

Utilize

Using or creating tools to communicate the results of data journalism tools is the most complex way. There are different tools, such as calculators, which are used often. However, there is also a wide range of other, more complex application since the amount of data increases from the publisher and the user. Once more, this part overlaps with the point personalization but can also work without personalization. The importance for subjects that are going to get utilized should be "durability" meaning a topic that is not trending but will always have relevance, such as unemployment. These tools should be kept updated and shared amongst the audience. In addition, live blogging and supportive visualization of exceptional or interesting points add value to the overall storytelling. Although, creating a utility or application has many advantages, it is still costly. As a result, of competition and growing standardization (usage of templates) the cost will probably decrease in the near future and news organizations can make use of such tools more often.

Summary

The essential idea of the visual data presentation is, as Alberto Cairo (2013) stated: "*Information graphics should be designed as if they were tools. As with every tool, they* *have to be functional*". This means that the visual shape and graphical form of the chosen data should depend and be constrained by its function.

2.2.5. Advantages of visualization

Due to the human evolution, the benefits of using data visualization or other visual stimuli are huge. The reason is that the human communication started to evolve between 100.000 and 30.000 years ago, while the written language has developed in the last 6.000 to 5.000 years (Larkin, 1999; Olson, 2014). Therefore, the reason that images are like shortcuts to the brain lies in the evolutionary and very visual past. For journalism in general, static and interactive graphics play different roles and can support it in many ways. They can help to identify reasons and questions during the reporting phase and it can also show errors within the data or a good story. Furthermore, visualizations can illustrate key points in a compelling and appealing way and simplify the written story, by removing complicated technical information from it. Especially, when interactive visualizations are used, it allows exploration and provides transparency about the background of the reporting. So, the benefits of information graphics or data visualization are according to Alberto Cairo (2013) that they represent tools, which make it possible to communicate, understand and analyze in a visual way. That means, visualizations make it easier to understand information, help to identify and reveal trends or information, which would otherwise be less evident. Cairo also pointed out another component that is an advantage of visualization, stating that, "*The brain doesn't just process information that comes though the eyes. It also creates mental visual images that allow us to reason and plan actions that facilitate survival*" (Cairo, 2013, p.XVI). Due to that, visualization of data makes it easier to absorb the information, because it allows the user to view and understand big amounts of abstract data and information. Moreover, it enables the reader to see connections between multi-dimensional datasets, and offers a way to interpret the data through graphical representations. As a result, of the creation of a mental visual image, it is also easier to retrieve the information from the visualization at a later stage. Additionally, the graphical data representations make it easier to discover relationships and patterns between the different parts and components. They also provide trends by showing historical data, which can be compared to current performances or situation and make it possible to forecast the future.

A specialized form are interactive visualizations because they provide a way to break up the data in various components, and see in how far it affects an individual or for example a specific region and how it changed throughout the history. By using interactive tools and viewing still visualizations the reader gains a traceable insight and interacts directly with the data, which cannot be done by just viewing a chart or table. As, Cairo (2013, p.XV) wrote: "*Everything our senses gather is transformed, deep inside our minds, into simple, manageable representations, or symbols*". These symbols can be a letter or a visual image and as displayed in fig. 8 verbal or textual language is more complicated to process than a visualization. Because, each letter in a word represents a symbol that needs to be decoded and put into context. That means, while the brain processes data from an image all at once, the textual data needs a linear process to understand the message.

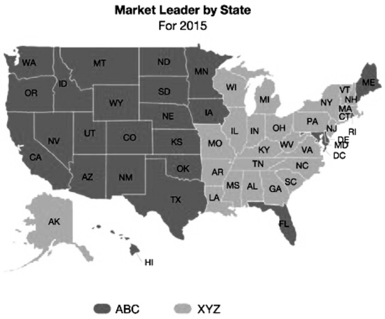
Summary

The use of images or visuals is by far more efficient to be processed and transmitted to the brain (Walter & Gioglio, 2015). That is due to the human evolution in which humans responded to visual information far before adapting to read text. In fact, visualizations and especially data visualizations provide helpful support in every stage of the data journalistic process and help to engage the audience. If the visualization is done well it makes the data accessible and understandable in an appealing way or like Priya Kumar (2015) summarized, "*Visualization is simply a tool to help traverse the gap between data and knowledge.*"

2.2.6. Problems of data visualization

Besides, making mistakes while researching data for a story, there are also mistakes that can be made when visualizing the information. For instance, providing incomplete data as displayed in fig. 13. The graphic shows that the company *ABC* has a higher market share regarding federal states but this does not mean that they reach the majority of the people. Hence, the graphic needs to present and display the numbers of the market shares. This is just one example that a data visualization can be misleading the audience, giving them a wrong impression of the relation of data, numbers and information that were gathered.

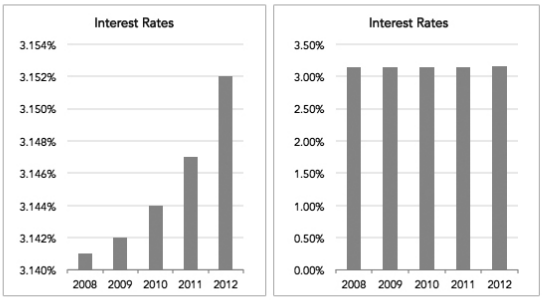
Fig. 13 - Incomplete data



In 2014, Ravi Parikh, cofounder of *Heap* published an article *How to Lie with Data Visualization*, which is closely linked to this topic. In the article, he emphasizes the simplicity to mislead the audience with rather basic visualizations, by naming and describing three of the most common ways. Additionally, he gave real-world examples to each misleading way, which are displayed alongside the descriptions in the following. Additionally, the MOOC *Math for Journalists* (Rodrigues, 2015) dealt with the problem of misleading statistics. Both sources agreed that the three most common mistakes or misleading ways are: "*truncated y-axis*", "*cumulative graphs*" and "*ignoring conventions*" (Parikh, 2014; Rodrigues, 2015).

Truncated y-axis

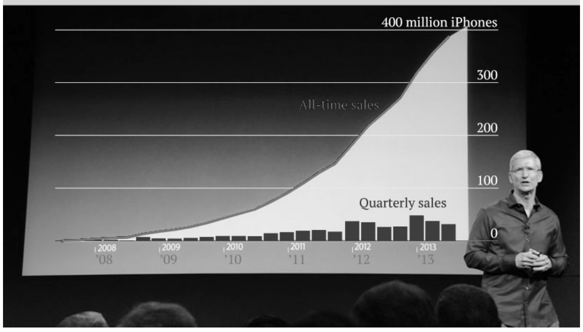
Fig. 14 - Same data, differernt y-axis



Describes the point of messing with the range of the y-axis, which usually ranges from zero to a specific maximum depending on the range of data. However, sometimes changing the range helps to show differences, but also makes differences look larger than they actually are as presented in fig. 14.

Cumulative graphs

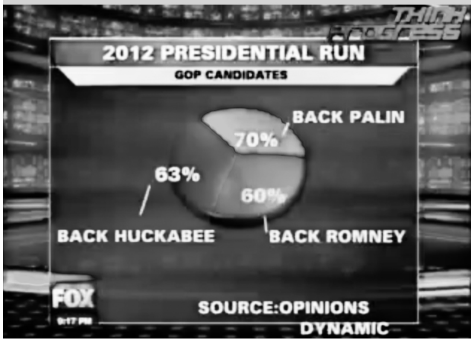
Fig. 15 - Cumulative graphs



This type of graph is mostly used to show, for example revenue or downloads. In the case of revenue, cumulative graphs of the total revenue earned to date, can be used rather than a quarterly revenue as in the example of iPhone sales (fig. 15). Referring to the cumulative graph, sales are going up but the added column graph shows a different picture. Nevertheless, the cumulative graph is misleading, since it is not immediately obvious that the quarterly sales of the product have declined over the last three quarters of 2013.

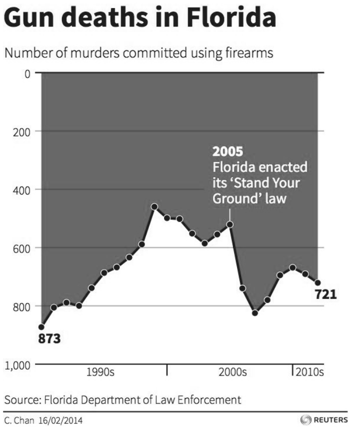
Ignoring conventions

Fig. 16 - Ignoring conventions - misleading pie chart



Another tactic to create misleading data visualizations is to break norms. Parikh gives the example of a pie chart, which does not represent parts of a whole as displayed in fig. 16. Another example of ignoring conventions is fig. 17 "*Gun deaths in Florida*", which shows an increase but due to turning the y-axis upside down gives the impression, as if the amount of gun deaths is decreasing. This examples, represents a clear violation of the commonly known convention, that the values on y-axis increase when moving up the axis. The final example is the visualization of *Top 10 Films by Worldwide Grosses* in fig. 18. Of course, a bubble chart represents a useful tool, to display three-dimensional data in only two dimensions. But, this leads to problem because a third factor plays a role - quantity. In the example are several mistakes but the focus for this example is put on the size of bubbles. They show the improper use of bubble sizes, which are not proportional to each other.

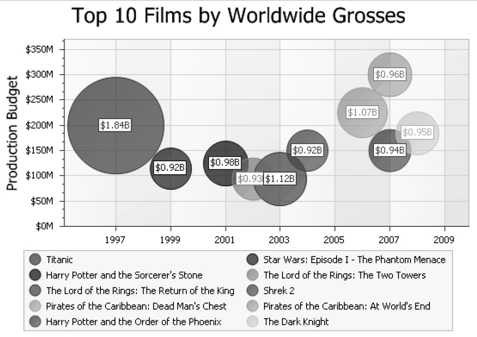
Fig. 17 - Ignoring conventions



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

Fig. 18 - Ignoring conventions - bubble chart



All of the examples given are also supported by Darrell Huff, author of the book "How to Lie with Statistics", in which he adds the component "visual impression" (Huff, 1993 p.71-73) when talking about the problem of using proportions in a volume based geometrical field.

Summary

To sum it up, there are many mistakes that can be made visualizing data but it should be remembered that data visualizations are supposed to make it easier to interpret data and further not to imply an opinion or feeling by breaking general rules. As a result, a journalist should have the knowledge of these conventions.

2.3. Dark Horse Vietnam

After exploring and reading through many similar approaches to data journalism and the concepts of the data visualization, they needed to be tested. This chapter presents the test of the theoretical workflow and evaluates the model.

2.3.1. Description

The main task of the project was to make use of the methods and ideas of open data journalism, in order to create a story with open, available data. Especially, with the background that the project marks my first data story including visualization that support and display the researched data, it can also be seen as a process of learning and studying. In order to, start the process it was suggested by data journalists to pick up and choose a topic that one genuinely wants to work on. Starting from there, the hacking and scraping of data happens until the project is working. Due to that, and the massive amount of data, which is provided by the Internet, the idea was to produce an open data story by making use of sources such as the *UN*, *Govdata* or *Eurostat*. So, using different sources, taking the workflows and the model of the *inverted pyramid* into consideration the article *Dark Horse Vietnam* was planned and inspired by the style of stories published on *The Guardian Datablog* and finally put into practice.

Dark Horse Vietnam

The motivation behind the article was a strong personal relation and my affection for coffee and the historical political relations between the former German Democratic Republic and Socialist Republic of Vietnam.

The main goal was to portray Vietnam’s coffee industry and the process of becoming the second biggest exporter of coffee in the world. Therefore, the story covers different aspects to give an insight into Vietnams coffee trade from the historical development, over the comparison with its competitors, to the global export and trade flow from Vietnam. The process and ideas behind the article are further explained in the next chapter while the article *Dark Horse Vietnam* can be found in the *Appendix 7*.

2.3.2. Process

The challenge was to turn something “off-topic” for most of the people, into a relevant and interesting story to pass on knowledge by giving them facts about the development and the current state using visual and narrative elements. The main idea of this project was to make use of the models and workflows explained above in the theoretical framework. In particular, Paul Bradshaw’s idea of *the inverted pyramid of data journalism* was to be tested and refined for smaller data stories. Furthermore, Edward R. Tufte’s and Alberto Cairo’s design suggestions were considered and applied to the visual components of data journalistic article and visualization. In order to understand specific items that are talked about in this chapter and to follow the workflow, see the article (*Appendix 7*) and take it by the hand.

The task was to go further than a short, visual data story by providing more details and a narrative, which give additional information that support, the presented graphics. The core of the data story is to show and tell the story of coffee from Vietnam, its global connection and the development of the coffee production of one of the most important coffee exporters.

The starting point of finding an appropriate topic that offers a variety of data but is also suitable to work on as a “first” data story took around two weeks. This general research, resulted in the realization that three of the four topics are very complex and will be worked on and finalized after this thesis is finished, leaving the very suitable topic of Vietnam and its strong appearance on the coffee market. This exceptional market position caught my interest because whenever people think of coffee, they think of Brazil, Guatemala or Ethiopia but not Vietnam.

After choosing this topic every step was documented and recorded, which became crucial when the data was gathered and changed. Furthermore, a dictionary for the different terms was made. The next step was to get further knowledge about understanding coffee trade and the coffee industry in Vietnam to think of possible questions to answer with data and follow throughout the research. This step inevitably leads to and represents Bradshaw’s first step - *Compile*.

2.3.2.1. Information and Data

Compile

Having found a topic and knowing that open data is available on this topic, which can be used in order to proceed, more accessible resources were rendered. The process of data gathering and compilation is in Bradshaw's point of view the fundamental part that defines the act of data journalism. He mentions two ways to begin a story, which can be a question that needs data to be answered or a dataset that needs questioning.

This project, started with the general question: "*Where does my daily coffee and 'fuel' for my master thesis come from?*". It did not take long to find out which countries exported the biggest amounts of coffee but the results were not what as expected. Especially, the second rank of Vietnam was a surprise and the main question was raised: "*How can Vietnam be the second biggest coffee exporter?*"

In the following, the data was scraped from the *UN Comtrade Database*, *International Coffee Organization (ICO)*, *Statista* and the *United States Department of Agriculture - Foreign Agricultural Service*. In order to, scrape the right data the correct commodity code needed to be investigated, which is 901 and involves coffee, coffee husks and skins and coffee substitutes. Prior to scraping the data, the sources and how they get the data was verified. With the exception of the *ICO* it was a simple and transparent task. The data of the *ICO* was not complete, so the question was raised for the reason. In reply to this question, it turned out that the data of the *ICO* is depending on the transparency of their member states.

Furthermore, some of the sources provided the data in *.XLS* or *.CSV* formats, while others provided *PDF* formats. Due to that, some tools were needed to make the data accessible. One was an *optical character recognition* (OCR) tool, in order to transform the *PDF* and access the data. Another one was a web-scraper which makes use of the *API* of a website to collect the data directly form the website. Further tools, were cloud-based programs to save the gathered "messy" data, in order to access it at all time with all devices. Consequently, the partly "untidy" data of the *PDF* and web scraping was put in an *Excel* and *Numbers* spreadsheet to start cleaning up the data. At that stage eleven different spreadsheets with raw data were put together.

Clean

The most essential part of the cleaning process was to isolate the data and adjustments of measurements because the common unit to measure the amount coffee is in 60 kg bags. Due to that, the value was adjusted to tons, to make it more understandable to the audience. The cleaning and adjustments were done using *Excel* and *Numbers*, since the datasets cannot be considered big data and therefore did not need other tools.

Also, the datasets itself were well structured and duplicates did not exist. Nevertheless, the process was time-consuming because empty cells had to be deleted and punctuation or formatting of the tables needed to be adjusted. Moreover, countries that do not exist anymore were deleted.

All in all, the readability of the spreadsheets was improved, to make it accessible for the audience but also to copy and use them in the "rudimental" spreadsheet of *Adobe Illustrator* during the visualization process. Consequently, each table existed in three forms, (1) raw original version, (2) cleaned version and (3) simplified version for the visualization.

Context

After cleaning and adjusting the data, data of an undefined area (Areas, NES) appeared when looking at Vietnam's exports of coffee in 2013. A smaller number could have been caused by human error but this item represented the second highest amount, which consequently raised the question of: *Who is behind this data?*

In first place, the outcome was unsatisfying, because the UN referred to the fact that they did not get any data on these exports and "*The reporting country does not send us the details of the trading partner in these specific cases*" (United Nations, 2010). But it was also mentioned that a human error or error by the reporting country can be a reason. Consequently, the Vietnamese customs was consulted, which could not provide any data on the exports of coffee for 2013. However, in my opinion this is a human error and the decimal place was set wrong, because this number does not appear in the years before and no development of this item can be recognized. This number also does not appear in any other statistics of, for example the *ICO*. Furthermore, it was important to check and know about the definitions and parameters of the *ICO* and how they gather the numbers, in order to consider or ask for different data. To important points were the facts, that the *ICO* calculates consumption as disappearance: imports minus re-exports plus or minus stock changes. In addition, the numbers provided were totals and usually did not add up owing to rounding. Both points did not affect the tables and were left aside. Nevertheless, it was important to question the data before continuing to work with it.

At this stage it was very helpful to have these questions noted down as guideline, focus and filter which data is important for the story. These questions were: *How developed Vietnams coffee industry? What is the current state of the production and trade of Vietnamese coffee?*

Combine

In the final stage, before thinking about how to communicate the numbers and story, the single spreadsheets were put together in one file. The combination was particularly helpful to look at the exports and imports in order to see if there is a big divergence in numbers, which might be another interesting story to look at and investigate. In this case, the numbers were within the scope and first ideas to communicate the story were collected.

Communicate

After all of the stages of Bradshaw's *inverted pyramid* were successfully mastered, the next step was to communicate the findings and story. For this, Bradshaw's second part of the *inverted pyramid*, *6 ways of communicating data journalism* were taken into consideration and are explained in the following parts.

2.3.2.2. Communicate

The model gives six approaches to communicate the gathered data. Those were evaluated if they can be used for this project and if they are appropriate to communicate the story. The idea was to make an easy to consume and appealing graphic with text and figures that makes use of most of the elements mentioned before without overloading the page.

The project is designed as a still image for print or a single page story, which does not have any interactive elements. This limitation results in a lack of some of the following points such as the utilization. However, in the following each type of communication is evaluated in connection to the project.

Visualization

Having the major advantage of communicating effectively was the reason to make use of this type. It was the most suitable way to visualize the numbers in pie and column charts. In addition, two maps were produced and used to provide an appealing way to show the global coffee consumption and coffee trade between Vietnam and the main importers of their coffee. The data visualizations were then visually connected to the coffee theme.

Narration

As mentioned before data stories do not need to include visual support, they can also provide data in text form by bringing down amounts to quantities that the mind can manage. The story also contains narrative elements, like most of the data stories on *The Guardian* but resulting the use of visualization, the narrative is less dominant. It is used to point out "highlights" to the audience. Nevertheless, the goal was to transform the complex data into understandable, concise and informal text.

Social communication

The story provides the possibility to communicate it on social networks and maybe even find an answer to the misleading numbers of the *Area NES* that was described before. As a matter fact, this story was only produced for this thesis, in order to test Bradshaw's model. Due to that, the social communication will be considered for future projects, in order to gather data and pass on findings.

Humanize

The approach to humanize and personalize the story was made in the narrative by addressing not only coffee drinker but also passionate coffee drinker that see their coffee as fuel for their everyday life.

Personalize

This point is strongly linked to interactivity, which cannot be provided due to its limitation of a still image. Nevertheless, the idea to provide the data sets and graphs with the data story is the analog, print form of interactivity by letting the eye and mind explore the data visualizations and find the information that is relevant to them.

Utilize

Another result of the print limitation is this point. Of course, there are possibilities to link data visualizations to an online application or utilities, which show the latest updates on an event such as the latest traffic alerts.

Summary

The article shows how supportive and helpful visualizations are to display numbers and data in an appealing way. They help the reader to explore the data, such as who the biggest consumers of coffee are. Furthermore, it offers a chance to dig deeper into the data by accessing the data, which lets the reader interact with the data.

The work itself showed that the majority of work is data cleaning (Patil, 2012) and was facilitated by the use of "clean" data by the mentioned sources. As a result, the work done up front to clean the data repaid over the course of the realization of the article (Patil, 2012).

Part II

Data stories are produced to communicate the findings, in order to engage the reader and attract especially the younger audience, because "[a] *report unread or not understood is a report wasted*" (Meyer, 2002, p.7). Accordingly to promote a data story, in particular to the Millennials, the use of social networks is unavoidable.

2.4. Millennials

Millennials or the Generation Y play an important role in relation to the case study, since they are attracted to social networks and represent the main user group. Furthermore, they need to be characterized to find out which and how their characteristics influence data journalism and if they could be attracted by its principles. This part gives an insight into this generation and describes their characteristics and trademarks.

The Generation Y is shaped by dramatic events, changing economic situations and a remarkable technical revolution all around the globe. Referring to Smola and Sutton (2002), a Millennial is a person born between 1979 and 1994 and represents a person of the first generation to come of age on the new millennium (Keeter & Taylor, 2009). The generation has already been characterized in a number of ways, for example in a survey regarding the appeal of brands for marketing reasons in which Millennials should reflect their personality (Barton, Fromm & Egan, 2012). According to this survey, Millennials described themselves as having unique personality traits, such as being technologically savvy, young, modern, risk taking, rebellious, hip, and funny or humorous (Boston Consulting Group, 2013). However, popular perceptions of Millennials are mostly negative and imply a very self-confident and self-absorbed personality, which is also described in the Pew Research Center report (2007). Further, they are commonly described as a social generation, which they express in real life and online. Moreover, this generation socializes while consuming different products and services. The socialization process happens as they review, post, update and share on different platforms like *Tumblr*, *YouTube* or *Wikipedia*, which reflects their eagerness to connect, collaborate and cooperate with each other and others. In other words, Millennials have a *we-can-fix-it-together* mindset (Solomon, 2015). In like manner, they also want to get along with everybody, which is a new phenomena concerning the authors van den Bergh and Behrer (2011, p.3), who stated that "[c]*ontrary to previous generations, Gen Yers were brought up in an atmosphere of equal relationships and co-decision-making (...)*". As a matter of fact, Millennials are motivated to collaborate with businesses, brands or news media as long as they believe their opinion matters to the company. Another characteristic of the Generation Y is their hunger for adventure and challenges that appear on their way even on an everyday base (Pew Research Center, 2007). The report, *Millennials - A Portrait of Generation Next - Confident Connected. Open to Change* (Pew Research Center, 2007) also describes them as generally happy, optimistic about their future and confident team player.

When looking at the political activity of Millennials, the Generation Y is interested in politics and keeps up with topics such as national affairs (Pew Research Center, 2007). But, referring to a poll from Harvard's Institute of Politics (Harvard University, 2015; Sky News, 2014), only 21% of them consider themselves as politically engaged. Furthermore, new political topics appeared amongst Millennials, such as "ethics" or "consumption". Due to that it can be said, they defined and are defining politics in a new way, which is different to their previous generation that fought for the change of the social order. Consequently, the new lifestyle of the Millennials leads to social changes, which makes them "*secret revolutionaries*" (Hurrelmann & Albrecht, 2014), because they change the traditional patterns of life. These changes also apply to the consumption of news and journalism.

Summary

The Millennials or Generation Y represents currently the youngest group of age. Since they are the generation that will influence the "world" regarding business, technology or media in the future, their characteristics and trademarks are widely discussed. However, two significant characteristics of this cohort, which are of interest for this thesis, are the connectivity through social networks and the fact that the generation scrutinizes norms, standards and values.

2.5. Social Network

In order to engage the audience, it is important to know who they are and a platform to communicate with them. In connection to data journalism, this means that the audience can access and discuss other important numbers, which have not been discovered by the journalist or did not fit with the data story the journalist aimed to tell. Since, Twitter represents a social network, this term needs to be clarified to give a basic understanding of its characteristics.

2.5.1. Terminology

The phenomena of social networks and media is not new because people have been providing information, recommendations and opinions to people they know in a two-way conversation before. In recent history, this happened using channels such as letters, telephones or more recent face-to-face discussions via video chat.

The term social network as it is used in this thesis, refers to an online community of individuals who create, share and consume content and information and/or cooperate on joint activities. Furthermore, it is considered a recent phenomenon that is experiencing a rising popularity over the last decade. The characteristics that distinguish the communication on "modern" social networks from other types of social interactions are referring to Nick Smith and Robert Wollan (2011) that they:

- Enable one-to-many or many-to-one conversations (peer-to-peer dialogues)
- Feature content created and posted by consumers of that content
- Are easy to use
- Are highly accessible, highly scalable and operate in real time (everyone + everywhere + every time)
- Entirely public and transparent

Antony Mayfield (2008, p.5), consultant on media strategy, social media and digital literacy defines social media by its characteristics, which are:

1. Participation – contribution and feedback
2. Openness – accessible to everybody
3. Conversation – two-way-communication
4. Community – share common interests
5. Connectedness – ability to link to other resources

To bring it to a point, "*(...) social media enables the swift and easy development, creation, dissemination, and consumption of information and entertainment by both organization and individuals*" (Smith & Wollan, 2010, pp.XI-XII), which is using social networks as its platform to interact. Therefore, the strength and advantage of social media and networks is the speed at which news, information are spread and the social component of building up relations amongst the user.

Summary

The main idea of social networks is the networking character, which is used to communicate media by status updates or posts in form of for example text, photos or videos. Common platforms are Twitter, *Facebook* or *Google Plus*, while the focus of this thesis is put on Twitter.

2.5.2. Twitter

Twitter is real-time digital microblogging platform, which enables users to spread and read 140-character messages. After the social networking platform was launched on the 15th of July 2006, it rapidly grew a huge community and gained worldwide popularity. By now it became one of the ten most popular websites referring to Alexa (2015) regarding its traffic. As of May 2015 they have 302 million active users, who are altogether sending out 500 million tweets per day with a continues growth (Twitter, 2015). Subsequently, this number gives an impression of the gigantic volume of data, which is produced. Accordingly, it makes Twitter a rich virtual source for information in sensing public opinion and discovering trends and development for political or other initiatives. In fact, these trends and trending topics are gaining importance for measuring the public opinion. Moreover, Twitter also became an important tool for companies to provide and spread information about their products and communicate this way with their customers. By that, they make use of Twitter as a marketing tool, and further use it to do research on the market to develop new products and future strategies.

Of course journalism also makes use of the potential of Twitter. Due to the short messages, current events can be found, reported and covered faster than professional made news stories. For example, the coverage of the *Aurora shooting* in 2012 by reporters of the *Denver Post* with immediate updates of only minutes in between the tweets (The Pulitzer Prizes, 2013). This reporting and coverage of the event was also awarded with the Pulitzer Price in Breaking News Reporting in 2013. Another example of making use of Twitter as a source, like the model that was made to show the Ebola spread in order to predict the next outbreaks or the Twitter Flu tracker. It is important to point out that due to the directness of the messages via Twitter unverified false reports might be multiplied and communicated.

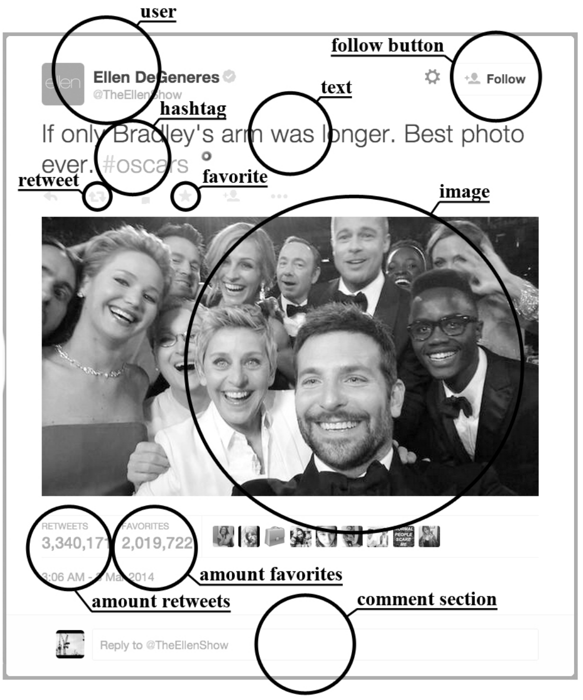
However, the microblog offers a platform to present different aspects from the life of the users or opinion on specific topics. These posts can then be commented or discussed by other registered readers. Consequently, tweets can be used to exchange information, thoughts or experience or other forms of communication. The content of tweets can be categorized in six main areas, referring to a study done by Pear Analytics in 2009 and quoted by Smith and Wollan (2011, pp.XI-XV). These six areas and their rounded shares amongst the tweets are:

- Pointless babble (40%)
- Conversational (38%)
- Pass along value (9%)
- Self-promotion (6%)
- Spam (4%)
- News (4%)

In order to get an impression of the general look, with the different components fig. 19 shows the anatomy of a tweet. The figure displays the most retweeted and shared tweet until now (June 2015), Ellen DeGeneres Oscar selfie (Twitter, 2014) from the Academy Awards in 2014.

Fig. 19 - Anatomy of a tweet

Schulze



Ellen DeGeneres
@TheEllenShow

If only Bradley's arm was longer. Best photo ever. #oscars

RETWEETS 3,340,17
FAVORITES 2,019,722
:06 AM - Mar 2014

Reply to @TheEllenShow

In the upcoming part, Twitters most relevant features for the case study are described and explained. Additionally, a glossary (Twitter Help Center, 2014) of the used vocabulary and terminology can be found in the *Appendix 8* for further explanations of the used terms and processes.

Tweets

A tweet describes a post which is made using Twitter. As mentioned in the introduction of this chapter, every tweet has a maximum of 140 unicode-characters and is public to everybody including not registered user. The tweet can include hashtags, links, other user profiles, images and locations.

Follow

By following a user or account one keeps track and stays updated by "subscribing" the tweets of a specific user, which are then shown in the timeline of the follower (person that subscribed the account).

Retweet

Who ever wants to share a tweet with followers can retweet with the integrated function. Tweets can also be quoted and commented. By retweeting a post the user usually passes along what they consider valuable news or other discoveries. Referring to Kwak, retweeting grows the voice of common people louder, while hundred thousands can listen (Kwak et al., 2010).

Favorite

The user has the possibility to like a tweet by marking it favorite. This happens by clicking the star icon. All of the favorite tweets can be found on the profile page of the user. The motivation to "Favorite" a tweet is different but it always implies that they like, agreeing or supporting with the content of the post.

Summary

Twitter is a microblog and represents a social network, which was used in the past and is still up until now to follow and cover stories and breaking news almost in real time. The users show and express their interest in stories or posts by retweeting (sharing) a tweet or marking them as favorite (liking), besides commenting. Furthermore, Twitter represents a platform of interest, due to demographic reasons and its global interconnection. Because of that it is also the source of the case study about the use of data visualization.

2.5.3. News consumption – Social networks

The fact that the Internet and in particular social networks provide a many-to-many communication makes it an appropriate medium to transfer and spread news. The fact tank *Pew Research Center* focused on the way news are spread and how the different channels relate to each other and examined the role of news on social networks in 2013 (Mitchell) and repeated the study in 2015 (Mitchell), with the result that the role of news on Twitter and *Facebook* is evolving.

The report shows that in the U.S., Twitter and *Facebook* are amongst the most relevant platforms for news consumption and users make use of these two networks in order to comply with their information needs. Moreover, 63% of these users get their news on those sites, which is a higher number in users than in 2013 (47% for *Facebook* and 52% for Twitter). This increase of users who get news on either Twitter or Facebook can also be observed throughout every demographical group. Nevertheless, the general usage of Twitter amongst U.S. adults (17%) is still lower than *Facebook* (66%). Another interesting point is the news habits on those sites, because particularly Twitter represents the main source for its users (59%) to keep updated about events as they are happening. In comparison only 31% of the *Facebook* user consider the site to follow breaking news. Amongst others, a relevant fact is that both platforms represent secondary sources for news but that especially the Millenials rely more on these networks for news (49% each on Twitter and *Facebook*). An additional finding of the survey was that Twitter news users are more likely than *Facebook* users to report seeing news about: "*national government and politics (72% vs. 61%), international affairs (63% vs. 51%), business (55% vs. 42%) and sports (70% vs. 55%)*" (Mitchell, 2015, p.3).

Ultimately, social media news are not only about consumption, it is also about engagement. Increasingly, the consumers or readers are contributing to the reporting with photos, videos or posts, which is a result of the growth of mobile devices.

Summary

The consumption of news on social networks has changed and especially Twitter and *Facebook* are important platforms for their users to get news. Both of these sites increasingly serve as source for news, which makes it attractive to present breaking news but also to find stories and get live updates on happenings.

2.5.4. Usage of Twitter

Around 646 million users (Statistic Brain, 2015) are registered at Twitter of which almost half (302 million) are active Twitter user. In order to understand the demographics and usage of Twitter, the *Pew Research Center* did a research about the demographics of social networking site users. Referring to the report, "[t]*he percentage of internet users who are on Twitter has doubled since November 2010 (...)*" (Duggan & Brenner, 2013). Of all Internet users, in 2012 16% (Ibid.) were using Twitter while the the latest standing from 2014 is at 23% (Duggan et al., 2015). The main group using Twitter is below the age of 50. Particularly, people aged 18 to 29 are very likely to use the social network, which also represents the group of interest - the Millennials. Of this group of users, half get their news from Twitter.

Millennials

The online statistics portal *Statista* (2015) published a chart based on a report by *comScore* in August 2014, which shows that Twitter is installed on the mobile devices of 23,8% of the U.S. American Millennials (18-34). It also states that besides *Facebook*, the social networking platforms *Snapchat* and *Instagram* are more popular amongst them than Twitter. These two platforms represent an exception, due to their characteristics and are very recent and young tech companies that could still show a hype that surrounds them. Certainly, both social networks are interesting platforms with potential regarding the presentation of data journalistic products. However, it is important to point out that the data provided by the *comScore* report is not about the actual usage of social media apps, but does indicate that *Snapchat* reaches a very high amount of young adults, especially in comparison to networks such as Twitter. In reference to the report *Profile of the Social Media News Consumer* by *the Pew Research Center*, which was taken into consideration before, "*Twitter news consumers are significantly younger than news consumers on Facebook, Google Plus and LinkedIn"* (Matsa & Mitchell, 2014). In numbers this means, 45% of the Twitter news consumer are between 18 and 29. Unfortunately, the study includes parts of the target group in a different age group, the 30 to 49 year olds (Mitchell & Guskin, 2013). Thus, it can only be estimated that the bigger share lies amongst the younger consumers, which would then give the Millennials a share of over half of the user on Twitter. Another fact about the user of Twitter, referring to the *Pew Research Center* study, is that the gender of the user is equally divided in 50% male and female. Due to that, Twitter represents a social network, which is most suitable when making a social network analysis on Millennials and their interests, because they represent a majority of the user group that produce and consume news on this platform.

Spread the word

To display how effective the network of Twitter works, Simon Rogers gave different examples in a presentation about *Twitter and data visualization* (2015). In the presentation, he chose two important beats: entertainment and sports. More specifically, he referred to the Oscars 2014 (#Oscars) and the transfers in football on the last before the transfer window closes, which is called "transfer deadline day" (#TransferDeadlineDay) and usually is the busiest day (Rogers, 2015). Another example given by him is related to *Typhoon Haiyan.* It shows how tweets and news are spread and how connected the Twitter community is. *Visualised: the world responds to Typhoon Haiyan on Twitter* (Belmonte, 2013) displays how Twitter was used by thousands of people to call for help after *Typhoon Haiyan* in 2013. In detail, an interactive map with every geo-tagged tweet that mentions the word “help”[6] in combination with key terms around the disaster was created. The result is the display of the massive global network.

All of the examples indicate how fast and easy news can be spread via Twitter. Furthermore, Shankland (2009) described this effect as *Twitter Effect*, which shows how tweets spread out like tree branches and reach a large amounts of users. As a result, Twitter represents a rich resource for studies and to gather information and data. Moreover, it is a suitable platform to spread news worldwide and reach in particular the younger audience, such as Generation Y and Z.

Summary

The global network of Twitter is an important tool for the audience and in particular the Millennials and their news consumption. As a result of its interconnectivity, the platform increases its meaningfulness because stories can be picked up easily through *hashtags* as they are tweeted and shared in the *twitterverse*.

2.6. Use of data visualization on Twitter (Case Study)

Social networking is a useful tool and could become an essential element for particularly data journalism. Due to that, it is interesting examine, to what extend news companies make use of social networks to publish their visual results of data stories but also in how far they are accepted by the audience. In the upcoming parts the approach, the way the study was obtained and the process of the research are described more in detail.

2.6.1. Explanation

Due to the instantaneous nature of social networks, journalists are not the first and fastest source to report about an event or happening anymore. This changes the role of the reporter, who can collect the information and data spread on social networks instead in order to show a development or explain the backgrounds of the events or issues.

Research period

The period of time to research the development is a result of implementation of the photo-sharing tool: *pic.Twitter.com*. Accordingly, the research period starts in July after

[6] In 22 different languages

the final launch of the image uploading service in June 2011 (Dorsey, 2011) and ends in December 2014.

Dataset

The whole set of data of the case study contains five different tables, which can be found in the *Appendix 10 to 14*. Each account provides an overview split in four parts, from 2011 to 2014. The gathered data and tables include the categories tweets, retweets, favorites and their averages.

Data gathering

As mentioned above all tweets showing a data visualization in relation to data stories were taken into account. This includes updated tweets such as maps, during breaking news for example, which were counted every time they had been tweeted since they showed new developments or information. With every relevant tweet, the amount of favorites and retweets were recorded, in order to conclude about the acceptance by the audience.

2.6.2. Analysis

In the following the collected data from the different news media accounts are presented and analyzed. Since, the numbers itself are very abstract and theoretical, every year provides a table to present the numbers in a comprehensible way. In order to make the analysis easier to read, the years are divided into quarters and summarized to point out the essential changes and developments related to the case study. Finally, a comparison between the different channels and their posts on a quantitative base are made and looked at.

The sum of all recorded, relevant tweets over the research period of almost four years is 2.963. The progress in posts over the years shows a huge growth in 2014 by over 450%, which means five and a half times more than in 2013. The same picture can be observed when looking at the feedback from the audience. The total of retweets and favorites is increasing bringing it to a total of more than half a million retweets (526.308) and 259.079 likes. Since the increase in retweets and favorites might be related to the increase in posts, the average retweet and favoriting per tweet was also taken into consideration and presents a rising trend as well. While every tweet was retweeted on an average of roughly 65 times in 2013, the average retweet of post in 2014 was 229, which is more than three times of 2013. As well as the average of retweets, the average favorite per tweet grew. More in particular, the average stars given to a tweet in 2013 was 28, which is four times less than the average like per tweet in 2014 (113).

So, the overall picture of the development brings up an ascending and positive trend in all researched categories. All of the accounts present an increase in usage of tweets with visual data journalistic products. Although, there is a slow growth to observe until 2013, numbers boosted in 2014. The only exception is the specialized account of *The Guardian* that already in 2012 shows a growth in numbers and dropped again in 2013, before rising again in 2014. This is related to a very interesting observation that was made during the research of this channel. The reason is that a lot of the data stories were posted using primarily photos or *The Guardian* logo as visual eye-catcher to attract the audience. Correspondingly, these posts could not be considered in the count and were not recorded in 2013, this changed in 2014.

To get a more detailed look at the numbers and the progression, the upcoming parts give a further insight. Moreover, a visualization of the shares amongst the years and the accounts can be found in *Appendix 9*.

2.6.2.1. Die Zeit

Die Zeit was first published on the 21st of February 1946 and is a German weekly newspaper. The web version Zeit Online works since 1996 and is also the provider of the Twitter account *@zeitonline*, which was examined in this case study.

@zeitonline

	Relevant Tweets	**Sum Retweets**	**Average Retweet**	**Sum Favorite**	**Average Favorite**
2011	2	30	15	20	10
2012	5	98	19,6	45	9
2013	25	629	25,16	291	11,64
2014	198	3.725	18,81	2.239	11,31
Total	230	4.482		2.595	

Table 1: Die Zeit - Overall

The channel is online since May 2005 and started tweeting on the 2nd of May. It has 686.210 followers (3. April 2015) and the overall amount of all tweets from the account is 71.412 of which only 230 are relevant for this case study and were considered. The development over the years is all in all positive and shows an increase throughout the categories, which implies a growing importance of tweets with a visual attraction in connection to data journalism. From year to year a

dramatic increase can be observed. As shown in table 1, every year the amount of post grew from 2011 to 2012 (150%), 400% from 2012 to 2013 and finally increased by 692% from 2013 to 2014. Also, when looking at the other numbers, like the growth of retweets and favorite, it can be said that *@zeitonline* is emphasizing its engagement in this field and gets rewarded by the users. However, the Twitter channel of Zeit Online represents among all researched Twitter accounts, the "weakest" one regarding the overall tweets, retweets and favorites. This has different reasons; one of them is the fact that all tweets are only posted in German and the account only has a small target group, which they can reach. The share amongst the four years looks as described in the following:

2011

Since Twitters photo-sharing tool was not in use in the first two quarters, there are no results, which could have been researched until July 2011. Nevertheless, the amount of all relevant posts in 2011 contains only two tweets. These were retweeted 30 times and marked as favorite 20 times and one each was found in November and December. To talk about a general development or analyze this year, the amount of tweets is too low. Due to that taking the average tweet rates into account is also unnecessary. This will be done in the following years with exception of 2012, which can also be considered a weak year as to see in the next part.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	0	0	0	0	0
Q2	0	0	0	0	0
Q3	0	0	0	0	0
Q4	2	30	15	20	10
2011	2	30	15	20	10

Table 2: Die Zeit - 2011

2012

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	1	3	3	0	0
Q2	0	0	0	0	0
Q3	0	0	0	0	0
Q4	4	95	23,75	45	11,25
2012	5	98	19,6	45	9

Table 3: Die Zeit - 2012

In 2012 a small increase on five overall retweets is to observe. One each can be found in January and November, with the tweet in January being retweeted three times and not marked as Favorite once and the November tweet being retweeted nine times and favorited seven times. December provides 60% of all the tweets (3), which were retweeted 86 times on an average of around 29 retweets per tweet. 38 users marked these three tweets as favorite, bringing it on an average of 12,7 markings per tweet.

As a result, the five tweets were retweeted 98 times (average of 19,6 retweets/tweet) and liked 45 times on an average of nine favorites per tweet. Nevertheless, as mentioned before the amount of tweets is also very rare in 2012 and an analysis of the quarterly development is rather relevant and not completely representative.

2013

This year provides an overall of 25 tweets with a sum of 629 retweets (25,2 retweets per tweet) and 291 favorites leading to average 11,6 likes per tweet. 2013 shows a more representative development than the last two years, so a more detailed analysis is possible. The development of the amount of relevant tweets is rather low compared to the other channels, but there is an overall positive progress throughout the year 2013 and an increase in amount of tweets, retweets and favorites (table 4) is to observe. The shares of the tweets throughout the year look like as described in the following.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	7	116	16,57	41	5,86
Q2	1	24	24	1	1
Q3	8	153	19,13	64	8
Q4	9	336	37,33	185	20,56
2013	25	629	25,16	291	11,64

Table 4: Die Zeit - 2013

Q1: This quarter shows an increase of 75% up to the sum of seven tweets. In addition, more retweets (116), which is a plus of 22,11% compared to Q4 in 2012 were recorded and a small decrease of 8,89% on 41 favorites can be observed.

Q2: Because of May and June, in which no relevant tweets were published, Q2 shows only one post and provides a big minus in comparison to Q1 and a decrease in absolute numbers in all categories as seen in the table 4.

Q3: However the second quarter seems more like an exception since a positive trend from Q1 towards the end of the year can be seen. So, the amount of tweets in Q3 rises up to 8, providing 153 retweets and 64 likes (favorites) by the Twitter users.

Q4: A positive trend can be seen and this quarter that provides one more tweet (9), which is a plus of 12,5%. Additionally, in Q4 the amount of retweets is more than twice as much compared to Q3 (336 retweets, increase of around 120%). Also the amount of favorite markings shows this increase (185) and is roughly three times more compared to the third quarter.

2014

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	45	1.199	26,64	628	13,96
Q2	43	1.359	31,60	785	18,26
Q3	54	481	8,91	331	6,13
Q4	56	686	12,25	495	8,84
2014	198	3.725	18,81	2.239	11,31

Table 5: Die Zeit - 2014

This year represents the highlight year in terms of relevant tweets, retweets and favorites. With an overall of 198 relevant posts, which were reposted 3.725 times (an average of almost 19 retweets per tweet) and liked 2.239 times, the year shows a very positive development. Especially, in second half of the year an increase in posts can be observed, which also shows a high amount of retweets and favorites.

Q1: The trend that can be observed during the second half of 2013 is continuing in the first quarter of 2014. The amount of tweets in Q1 boosts up by 400% on 45 tweets. The same development is seen in the retweets (1.199), which means a plus of around 257% and the favorites (628 with an increase of around 240% to Q4, 2013).

Q2: This quarter provided two tweets less than Q1 (43) but still increased the amount of retweets by around 13% on 1.359 and also the favorites by 25% on 785 likes. More than half of the post can be found in April (24) and also more than half of the retweets and favorites are found in April. Nevertheless, Q2 shows a positive trend overall, which continues in some categories in the third quarter.

Q3: Q3 shows an increase by more than a quarter, in comparison to Q2 in total numbers of posts, up to 54 relevant tweets. These were retweeted only 481 times, which is a decrease of almost two thirds (rounded 65%) in relation to the last quarter. A similar development can be observed in the likes of the tweets (331 favorites) that declined by more than half (rounded 58%), in comparison to Q2.

Q4: In the last quarter of the relevant research period an overall amount of 56 relevant tweets was recorded. This also marks the highest amount posted within the whole period of time that was researched and led once again to an increase of almost 4% compared to the Q3. Furthermore, the amount of reweets reached 686 (+42,6% to Q3) and the tweets were also liked almost 50% more (495 times) than in the last quarter.

2.6.2.2. The Guardian

The newspaper was first published in 1821, when it was called *Manchester Guardian* until 1959, after that it became *The Guardian*. It is a British daily newspaper and the owner of the Twitter-Account *@guardian*. Additionally to that account, *The Guardian* has a more specialized account *@GuardianData*, which was also taken into consideration and is the reason this section is separated in two separate parts.

@guardian

The Guardian joined Twitter in November 2009 but did not tweet until the 11th of December 2009. The amount of tweets until the 3rd of April 2015 is 134.862 and around 3.45 million people follow the channel. The account of *The Guardian* posted throughout the four years 435 relevant tweets, which got a feedback of 36.178 retweets and were favorited 16.063 times.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
2011	27	725	26,85	167	6,19
2012	60	2.535	42,25	1.362	22,7
2013	62	4.050	65,32	1.720	27,74
2014	286	28.868	100,94	12.814	44,8
Total	435	36.178		16.063	

Table 6: The Guardian - Overall

The development of the numbers is increasing from 2011 to 2014, experiencing a big boost from 2013 to 2014 as shown in table 6. As a result of this recorded data and the ascending trend in all categories, *The Guardian* and the audience imply a growing interest in consuming and posting data journalistic products. These posts are spread amongst the years as described next.

2011

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	0	0	0	0	0
Q2	0	0	0	0	0
Q3	9	273	30,33	48	5,33
Q4	18	452	25,11	119	6,61
2011	27	725	26,85	167	6,19

Table 7: The Guardian - 2011

In the first year that was researched, 27 tweets that are relevant to the study were recorded. Those were 725 times retweeted (average of 26,9 retweets per tweet) and marked as favorite exactly 167 times (6,2 markings/ tweet). This year provided a good opening, for the following years and shows that data journalistic products already found a niche to attract the audience at an early stage.

Q1/Q2: Resulting the late implementation of Twitters photo-sharing tool, no relevant posts were published during the first and the second quarter.

Q3: In the first quarter to be researched, nine relevant tweets were counted, which were retweeted 273 times and liked 48 times.

Q4: From the good start in the last quarter to Q4 the amount of tweets doubled up to 18 tweets. This increase also affected the amount of retweets (452) and favorites (119) positively.

2012

The 60 relevant tweets, 2.535 retweets and 1.362 favorites draw a positive development in all categories. From the first to the last quarter an ascending trend can be seen, with a small drop in the Q3. The exact numbers throughout the quarters are explained in the following.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	13	600	46,15	284	21,85
Q2	18	591	32,83	386	21,44
Q3	11	417	37,9	269	24,45
Q4	18	927	51,5	423	23,5
2012	60	2.535	42,25	1.362	22,7

Table 8: The Guardian - 2012

Q1: In comparison to the last quarter of 2011 the amount of tweets went down by 28% to 13 tweets. Although the amount of relevant posts fell, the feedback, in form of retweets (600) and favorites (284) has risen as shown in table 8. As a result, the average retweet per tweet is 46 and average favorite 22.

Q2: The overall amount of tweets in Q2 went up to 18 and so did the amount of likes (386). While the tweets were retweeted 591 times, a small decrease of 9 retweets can be observed.

Q3: The drop in numbers mentioned above in Q3 looks as the following. On one hand the amount of all categories decreased, tweets (11), retweets (417) and favorite (269). On the other hand, the averages of retweets and favorites show an increase compared to Q2. While in Q2 the average retweet per tweet was around 33 and 21 favorites per tweet, the third quarter brings up an average retweet of around 38 and around 25 favorites per tweet.

Q4: This quarter brought up the highest amount so far in retweets, which have more than doubled (927) and favorites (423) that went up by over 50%. Meanwhile, the provided amount of tweets (18) is the exact same as in Q2.

2013

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	13	742	57,08	284	21,85
Q2	7	224	32	83	11,86
Q3	10	502	50,2	224	22,4
Q4	32	2.582	80,69	1.129	35,28
2013	62	4.050	65,32	1.720	27,74

Table 9: The Guardian - 2013

The year 2013 provides two more relevant tweets (62) than 2012. But the retweets (4.050) and favorites (1.720) developed positively. In particular the amount of retweets increased by almost 60% compared to 2012. Looking at the overall development, an ascending trend from the beginning to the end of the year can be seen but again the numbers drop in the middle part of the year The average retweet in 2013 adds up to 65,3 and the average favoriting is 27,7 per tweet. The numbers of the categories are spread amongst the quarters as seen in table 9 and shown in the following.

Q1: The first quarter brought up 13 tweets, which were retweeted 742 times and liked 284 times. Once again the trend from the previous year's Q4 is a decreasing one. Even though the average retweet per tweet with around 57 is higher than in Q4, 2012.

Q2: The amount of tweets almost halved down to seven tweets in this quarter. Due to that also the retweets (224) and favorites (83) were affected and further the related average in both categories.

Q3: The numbers recover after the weak quarter before, providing 13 tweets that were retweeted 502 times and favorited 224 times. Nevertheless, the amounts are still below the result of Q1.

Q4: After the low in the two previous quarters, the amounts of all categories got a boost and the amount of tweets are more than twice as much as in Q1. The 32 tweets were shared 2.582 times and were 1.129 times marked with a star. This is an increase of numbers of over three times in these two categories from Q1 to Q4 (retweet: 248%, favorite: 298%) Moreover, this is the first time the two categories reach a score, above 2.000 or 1.000 hits.

2014

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	29	4.733	163,21	2134	73,59
Q2	79	7.568	95,8	3166	40,08
Q3	95	8.674	91,31	4086	43,01
Q4	83	7.893	95,1	3.428	41,3
2014	286	28868	100,94	12814	44,8

Table 10: The Guardian - 2014

The last year that was researched, shows an overall of 286 tweets and an ascending trend throughout the year as shown in table 10. Furthermore, the tweets were retweeted 28.868 times while liked with an overall of 12.814 stars, which led to an average of 101 retweets/tweet and 45 likes/tweet. The shares of the different categories amongst the quarters look like that:

Q1: As in the years before the first quarter presents in comparison to the previous quarter, a decrease in the amount of tweets (29). However, the other two categories (retweet: 4.733 and favorite: 2.134) show a massive increase by over 80%. The first quarter also shows the highest average of retweets (163 retweets/tweet) and favorites 74 stars/tweet) during the entire research of the account.

Q2: 172% more tweets than in the last quarter is the result of the recorded data for Q2. With 79 tweets, 7.568 shares and 3.166 favorites a massive increase in numbers can be seen. The retweets increased by almost 60%, while the favorites went up by almost 50%.

Q3: The highest scores throughout the whole research of this account were reached in Q3. A total of 95 relevant tweets were counted, with an overall of 8.674 retweets and 4.086 favorite markings.

Q4: In comparison to the previous quarter a decrease in numbers (tweet: 83, retweet: 7.893 and favorite: 3.428) can be observed. Nevertheless, the numbers still mark the second highest amount overall and tops the strong year, while the overall ascending trend continues.

@GuardianData

The account of *The Guardian's Datablog* is active since the 11^{th} of March, while joining in March 2009. So far they posted 6.986 tweets and have 54.729 followers (3. April 2015). The sum of all relevant tweets that were recorded from the account *@GuardianData* is 588 that were retweeted 10.396 times and liked (favorite) 3.947 times, which is a remarkable number especially when looking at the amount of follower of the channel has. Also the development of the numbers over the years show a steady rise in all categories (tweets, retweets, favorites) as presented in table 11, with the exception of 2013. Even though, the numbers are lower than in 2012 it is important to point out that in comparison to 2011 the overall amount of all recorded data is still higher. This leads to another interesting observation, the increase of retweets and clicks of the star button (favorite), which implies a gaining interest. From 2011 to 2014 both categories show a rising average of the retweets and favorite per tweet (2011: 5,7 retweets/tweet to 2014: 32,3 retweets/tweet and 2011: 3,2 favorites/tweet to 2014: 10,1 favorites/ tweet). The development cover the years can be seen in the table and are further described below.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	0	0	0	0	0
Q2	0	0	0	0	0
Q3	13	67	5,15	64	4,92
Q4	44	258	5,86	118	2,68
2011	57	325	5,7	182	3,19

Table 11: GuardianData - Overall

2011

The specialized account of *The Guardian* already provides a sum of 57 relevant tweets that were recorded. These were retweeted 325 times and favorited 182. Altogether, this year brought up usable data to work with and make a first analysis, which shows an ascending trend in the use and acceptance of data visualizations. Meanwhile, the shares amongst the quarters look like this:

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
2011	57	325	5,7	182	3,19
2012	217	1.732	7,98	964	4,44
2013	90	1.095	12,17	531	5,90
2014	224	7.244	32,34	2.270	10,13
Total	588	10.396	17,68	3.947	

Table 12: GuardianData - 2011

Q1/Q2: Due to the use of Twitters own image tool, which was mentioned before there is no data available for the first quarter and neither the second quarter.

Q3: This quarter provides the first relevant tweets, with the first "usable" post in August and a total of 13 tweets that were retweeted 67 times (average around 6 retweets per tweet) while being like 64 times (average of around 5 favorite markings per tweet).

Q4: Once again an ascending in the amount of relevant tweets up to 44 can be seen, which is three times more (rounded 239%) than in Q3. The same scenery can be observed when looking at the amount of retweets (258, increase by almost 290% compared to Q3) and the favorites. In comparison to the amount of tweets and retweets, the favorites "only" made it to escalate their amount by almost 85% up to a total of 118 markings.

2012

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	40	209	5,23	142	3,55
Q2	69	542	7,86	326	4,72
Q3	49	310	6,33	166	3,39
Q4	59	671	11,37	330	5,59
2012	217	1.732	7,98	964	4,44

Table 13: GuardianData - 2012

The trend that started to appear in the end of 2011 continues in 2012, where an overall amount of 217 relevant tweets were counted. Also the amount of reposts and likes increased massively and went up to a total of 1.732 retweets and 964 favorite markings. In particular, the development of retweets and favorites are impressive. Both categories provide the high scores in Q2 and Q4. To be more precise from Q2 to Q4 the amount doubled (retweet: 116%) or almost doubled (favorite: 99%) their value. In comparison, the rate of retweets was 7,9 retweets per tweet in Q2 while in the fourth quarter every tweet was retweeted at an average of 11,4 times. The same phenomenon is to observe when looking at the favorites. While the tweets in Q2 were liked on an average of 4,7 times, the users favorited a tweet on an average of 5,6 in Q4. With the increase of numbers in the fourth quarter, 2013 shows an overall positive trend in the development. Although, Q2 came along with a surprisingly high amount of researched data, the following two quarters brought up good numbers and presents an overall amount, which is still higher than the numbers of the first quarter.

Q1: Nevertheless, the first quarter of 2012 shows a loss of four tweets (-9% down to 40 tweets) and a decreasing amount of retweets (209, rounded 19%) compared to the last quarter of 2011. However the favorites on the other hand increased, by approximately by 20%, up to 142 likes.

Q2: This quarter marks the strongest part regarding tweets so far of *@GuardianData* and also throughout the whole year 2012. The three months provide in sum 69 relevant posts, which were retweeted 542 and favorited 326 times. As a result, of this high number an ascending of all collected data can be seen. In detail, the tweets increased by round 73%, the retweets by round 159% which is 2,5 times more than in Q1 and the favorites are more than twice as much as in the last quarter.

Q3: After the strong second quarter the amount of tweets goes down by 29% to 49 tweets and also retweets (310) and favorites (166) lose ground. However, the amount is still above Q1 and the trend experienced an exceptional high in Q2 but on the average is still ascending which is kept up in the next quarter.

Q4: The relevant tweets in Q4 include a total of 59 posts, which were retweeted 671 times and were liked 330 times. In comparison to Q3, this means an increase of the relevant post by approximately 20% is to observe.

2013

The positive trend that can be seen in the last two years comes to an end in 2013. The overall tweets only come up to 90 relevant posts. These were retweeted 1.095 times and marked as favorite 531 times. However, it is important to point out that the retweets per tweet (average of 20,4) are the highest average of retweets since data was recorded for this case study. The same observation can be made when looking at the amount of likes per tweet. With an average of 8,2 star markings per tweet this also is a high score up until this point. Overall, from the first to fourth quarter in 2013, a clear descending trend in numbers can be observed and every quarter shows a negative development in comparison to Q4 of 2012. However, the last quarter in 2013 shows a positive development in comparison to Q3 and displays a recovery, while marking the start of an ascending trend that continues in 2014.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	35	407	11,63	216	6,17
Q2	21	174	8,29	118	5,62
Q3	16	146	9,13	50	3,13
Q4	18	368	20,44	147	8,17
2013	90	1.095	12,17	531	5,9

Table 14: GuardianData - 2013

Q1: As already described the numbers of all categories dropped, which can also be seen in the first quarter. The overall number of the relevant tweets is 35 and is a decrease by around 41%, compared to Q4 2012. These posts were retweeted 407 times and favorited 216 times. Furthermore, the first quarter marks the highest recorded numbers throughout the whole year and the descending trend continues in Q2.

Q2: This quarter provides 40% less tweets (21) than the last quarter. The amount of retweet is 174, which is less than half (-57,3%) of the retweets in Q1. This development can also be seen when looking at the sum of favorites (118) and means a minus of 45,4%.

Q3: Also the third quarter joins the drop in numbers and falling trend that is described above. With only 16 relevant tweets, 146 retweets and 50 favorite markings the third quarter marks the lowest recorded numbers since Q3 in 2011, with the exception of the zero scores in Q1 and Q2 in the same year.

Q4: In this quarter the numbers recover slowly and an increase in numbers can be observed. A total of 18 relevant posts were recorded, which is an increase by 12,5% but compared to Q1 still shows a dropping trend in the amount tweets, retweets (368) and favorites (147).

2014

After a recession in the last year, the numbers slowly recover in 2014 and show a strong ascending over the four quarters. In sum, 224 relevant tweets were published and retweeted 7.244 times, which means an average of circa 32 retweets per tweet. A similar development shows the favorite category that reaches an overall amount of 2.270 likes. Looking at these numbers, already gives an impression of 2014 and the positive development in comparison to 2013 and the years before. After looking at the development throughout the four quarters in 2014 more in detail, a strong increase in numbers from the amount of tweets, over retweets and favorites can be seen. In Q3 and Q4 the sum of retweets reached more than 2.500 and a similar development can be observed in the other researched categories. All in all, from Q1 to Q4 an ascending trend can be recognized, which flattens slightly in the fourth quarter after a very strong increase in Q3. Nevertheless, the recorded data of the last quarter of the year is still very high within all categories and marks the highest numbers throughout the year. Moreover it shows a gain in acceptance with an average of 42 retweets per tweet and an average of

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	24	540	22,5	210	8,75
Q2	40	701	17,53	256	6,4
Q3	78	2.553	32,73	639	8,19
Q4	82	3.450	42,07	1.165	14,21
2014	224	7.244	32,34	2.270	10,13

Table 15: GuardianData - 2014

favorite markings of 14/tweet. In detail, the share amongst the quarters and development looks as the following:

Q1: In the first quarter of 2014, 24 relevant tweets were recorded which is an increase by one third from the last quarter 2013 to this quarter. Furthermore, the amount of retweets and favorites increased by more than 40% up to 540 retweets and 210 favorites.

Q2: The positive development continues during the second quarter and the numbers rise once again. So, the sum of all relevant tweets (40) is two thirds more than in Q1. A similar process is seen in the overall retweets (701) and favorites (256) in Q2. However, the escalation is not as high as the one observed at the tweets, which leads to a lower rate of average retweets (17,5 retweets/tweet) and average likes (6,4 markings/ tweet) than in Q1 (22,5 retweets/tweet and 8,75 favorite/tweet).

Q3: A big boost up can be seen in the recorded data in Q3 and in table 15, leading to an overall of 78 tweets. That is the highest amount of tweets so far and from Q2 to this quarter the relevant tweets almost doubled (95%) and also the amount of retweets (2.553) and favorites (639) increased massively. In case of the retweets, it is 2,5 times more than in the last quarter and the numbers in the category favorite increased by roughly 150%, which means it is more than twice of the likes than in Q2.

Q4: In the last researched quarter the data points out that the ascending trend continues throughout the year. After the strong the third quarter, this quarter tops all recorded numbers. The sum of all tweets (82), retweets (3.450) and favorite (1.165), all show the highest amount throughout the recorded years leaving a very positive picture of the use of data visualization of *@GuardianData*.

2.6.2.3. The New York Times

The New York Times is the U.S. American representative within the case study. The newspaper was first published on the 18th of September 1851 and is a daily newspaper. As *The Guardian*, *The New York Times* has an official Twitter account *@nytimes* and an account, more specialized in graphics and data visualizations called *@nytgraphics*.

@nytimes

The official account, *@nytimes* was opened in March 2007 and started to operate and tweet on the 5th of March 2007. Until the 3rd of April 2015, 175.428 tweets were posted on the channel and more than 16 million users followed the channel.

In sum, 1.319 relevant tweets were published and retweeted 447.816 times, while favorited 224.222 times. The recorded data of this channel represents the highest amounts of all researched accounts during the period of research. Meanwhile the trend of the development over the years is ascending and provides an increasing amount of numbers. Moreover, a big increase in numbers from 2013 to 2014 in every category was recorded. In detail this means that the amount of tweets are seven times higher in 2014 than in the previous year. Also the retweets are almost 23 times more than the retweets in 2013 and a similar development can be seen when looking at the favorites, which are 27 times higher in 2014 than in the year before. The positive development and growing acceptance of the consumers amongst the years is described in the upcoming part.

	Relavant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
2011	4	563	140,75	194	48,5
2012	20	2.047	102,35	601	30,05
2013	161	18.591	115,47	7.978	49,55
2014	1.134	426.615	376,2	215.449	189,99
Total	1.319	447.816		224.222	

Table 16: The New York Times - Overall

2011

The two quarters of the period of research provide an amount of four relevant tweets, of which three were found in Q3. Those three posts were retweeted 296 times and favorited 100 times. The fourth tweet was found in the fourth quarter and retweeted 267 times and favorited 94 times.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	0	0	0	0	0
Q2	0	0	0	0	0
Q3	3	296	98,67	100	33,33
Q4	1	267	267	94	94
2011	4	563	140,75	194	48,5

Table 17: The New York Times - 2011

The recorded numbers of tweets is too low to present a trend throughout the year, but the amount of retweets (563) and favorites (194) is the highest score compared to the recorded data of all the other accounts in 2011 and already shows a huge acceptance by the audience. The average retweet and favorite in 2011 is 141 retweets/tweet and 49 favorites/tweet, which mark the second (average retweet) and third (average favorite) highest average of the account.

2012

From 2011 to 2012 the overall amount of tweets went up to 20 posts. They attracted many users and all tweets were retweeted 2.047 times and favorited altogether 601 times. The year starts stagnant in the first half before the recorded numbers start raising and show and ascending trend towards the end of the year. Overall it can be said that the feedback in form of shares and likes is higher than the two categories amongst the other channels.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	1	125	125	65	65
Q2	1	37	37	13	13
Q3	8	660	82,5	229	28,625
Q4	10	1.225	122,50	294	29,4
2012	20	2.047	102,35	601	30,05

Table 18: The New York Times - 2012

Q1: As mentioned, the start is with one post a slow one. This one post has been reposted 125 times and was liked 65 times, which marks a very high amount of average feedback.

Q2: The described stagnant trend continues in the second quarter that also brought up one post. However this tweet did not get as much feedback, with 37 retweets and 13 stars as the post in the last quarter.

Q3: An increase in tweets, up to eight relevant posts, can be seen in Q3. Henceforth, also retweets (660) and favorites (229) were affected and increased, starting a positive trend towards Q4.

Q4: The trend continues in the last quarter of 2012 and in Q4 ten relevant tweets that were overall reposted 1.225 times and marked 294 times with a star, were recorded.

2013

The positive trend in numbers continues in 2013, with eight times more tweets (161), nine times more retweets (18.591) and 13 times more likes (7.978) than in all 2012. The development throughout the year is also positive and the rising trend continues as shown in table 19 and described in the next part.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	13	1.496	115,08	490	37,69
Q2	32	5.156	161,13	1.660	51,88
Q3	41	2.423	59,10	1.192	29,07
Q4	75	9.516	126,88	4.636	61,81
2013	161	18.591	115,47	7.978	49,55

Table 19: The New York Times - 2013

Q1: Compared to last year’s quarter, a plus of three tweets (13) can be recognized. Furthermore, reposts and likes have risen to 1.496 (retweets) and 490 (favorite), which continues the positive trend from the previous months.

Q2: In this quarter the relevant tweets jump up to 32 posts, which is more than twice as much as in the previous quarter. This development can also be seen in retweets (5.156) that provide more than three times of Q1’s shares and also the favorites show an increase of more than three times compared to Q1 with an overall of 1.660 star markings in this quarter.

Q3: The third quarter shows a growth in tweeted content by 28% up to 41 tweets, while registering a drop in retweets and favorites. The amount of reposts (2.423) is only half of Q2’s and also the likes went down by more than one-fourth to 1.192.

Q4: After the drop in the last quarter, 75 tweets were recorded in Q4, which is the highest amount until here. Also the feedback of these posts is the highest so far, while the retweets went up to 9.516, the favorites rose up to 4.636 clicks.

2014

The last year in the period of research, marks the highest scores among all recorded data of all accounts. The whole year provided an overall amount of 1.134 tweets, a sum of 426.615 retweets and 215.449 favorite markings. Meanwhile, the development of the numbers continues the positive trend observed in the previous years. Although, the last quarter shows a drop, compared to Q3. Nevertheless, the overall trend is ascending and the numbers show a steady increase in every category.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	104	17.243	165,80	11.323	108,88
Q2	248	99.535	401,35	60.715	244,82
Q3	434	137.256	316,26	81.679	188,20
Q4	348	172.581	495,92	61.732	177,39
2014	1.134	426.615	376,20	215.449	189,99

Table 20: The New York Times - 2014

Q1: In the first quarter 104 tweets can be found, which is an increase by almost 40% compared to the last quarter in 2013. During the quarter all of the tweets were retweeted 17.243 times (increase by more than 80%) and liked 11.323 times, which is almost two and a half times more than in Q4 2013.

Q2: With the amount of 248 relevant tweets the ascending trend is kept up. Correspondingly, the numbers of retweets (99.535, average of 401 retweets/tweet) and favorites (60.715, average of 245 markings/tweet) increased. This average in favorite markings per tweet also is the highest throughout the whole research.

Q3: These three months, mark the quarter with the highest amount of absolute numbers in tweets (434) and total of likes (81.679). The amount of retweets also continues to increase up to 137.256, which has not reached the maximum yet.

Q4: In this quarter, the highest total of retweets (172.581) by 348 posts was recorded (-20%). This leads to an average of 496 retweets per tweet, which marks an all time high throughout every account during the research period. However, those tweets were favorited 61.732 times, which is a loss of almost one quarter compared to Q3.

@nytgraphics

Meanwhile, the specialized channel run by *The New York Times* graphics department has currently (3. April 2015) 66.942 followers and published 1.434 tweets on their account. They joined in November 2009 and tweeted the first time on 26th of November 2009. The specialists of *The New York Times* tweeted an overall of 391 relevant tweets in the research period. These tweets provide a total of 27.436 retweets and were marked favorite 12.252 times. The development from 2011 to 2014 shows a steady growth in all three main categories as shown in table 21. This trend is particularly impressive when confronting the amount of three tweets from 2011 with 315 tweets in 2014 or their retweets. In fact, the amounts of tweets have increased within three years by 100 times, while the retweets are more than 300 times higher and the favorites went up by 560 times from 2011 to 2014. This development implies that the interest in such posts has increased over the years and the interest of the users, especially of the *@nytgraphics* account is growing. This is also indicated by the positive development of the average retweets and favorites rate per tweet. In detail, the tweets are spread amongst the years as seen in the following:

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
2011	3	82	27,33	21	7
2012	22	234	10,6	82	3,73
2013	51	750	14,7	389	7,6
2014	315	26.370	83,71	11.760	37,33
Total	391	27.436		12.252	

Table 21: New York Times Graphics - Overall

2011

Although, this is the specialized account of *The New York Times*, the overall amount of relevant posts in 2011 only includes three tweets. These were retweeted 82 times and favorited 21. Two of them were recorded in the third quarter (35 retweets, 14 favorites) and one in the last quarter. The sum of retweets increased compared to the third quarter to 47 retweets. While the reposts went up, the average amount of favorite stayed the same with 7, which is also the overall amount in the last quarter. With only three tweets this year, a solid and reliable analysis is impossible to make, even though the amount of retweets is interesting to mention since it is higher compared to the other accounts.

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	0	0	0	0	0
Q2	0	0	0	0	0
Q3	2	35	17,5	14	7
Q4	1	47	47	7	7
2011	3	82	27,33	21	7

Table 22: New York Times Graphics - 2011

2012

This year provides more relevant tweets (22) than 2011 that were retweeted 234 times, which adds up to an average of almost eleven retweets per tweet. Moreover, all tweets were marked as favorite 82 times. The development of the amount of relevant tweets from the beginning towards the end of the year is a very positive one and also the 234 retweets of all posts shows that the audience already had a high acceptance and affinity to this kind of journalism at an early stage. Furthermore it is important to mention that several tweets with a data journalistic background and data visualization were tweeted but not recorded. The reason for that is that they contained the logo of *The New York Times* instead the image of data visualizations. The mentioned 22 relevant tweets are shared amongst the quarters like this:

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	1	44	44	18	18
Q2	0	0	0	0	0
Q3	2	7	3,5	4	2
Q4	19	183	9,63	60	3,16
2012	22	234	10,64	82	3,73

Table 23: New York Times Graphics - 2012

Q1: The overall amount of tweets stayed unchanged from the last quarter of 2011 to this quarter. Again, only one relevant tweet was posted, which was retweeted 44 times and liked 18 times overall.

Q2: In this quarter, no tweet with relevant content was posted and therefore no data was fetched.

Q3: This looks different in Q3; the quarter provides an amount of two relevant tweets that were retweeted seven times and favorited four times. Once again no real development can be observed and as the amount of tweets is very low.

Q4: A new picture is drawn during this years fourth quarter, in which 19 tweets were posted that are relevant for this case study. Also the amount of retweets (183) and favorites (60) experiences a massive boost.

2013

After the weak last years, 2013 provides a good amount of relevant tweets with an overall of 51, which means it has more than doubled from last year to this year. Also the sum of retweets (750) and favorites (389) show a drastic increase. As already stated in Q2, this year marks the beginning of a positive trend towards the use of visual products such as data visualization within a tweet to attract the audience and gain more attention. The increase of attention by the readers indicates the growth in numbers of retweets and favorites. Overall the year leaves a positive impression, in particular when looking at the strong numbers in Q4. This positive trend and impression is supported by the relevant data published in 2014, because of the higher numbers this year gives the first real opportunity to look at the development amongst the different quarters:

	Relevant Tweets	Sum Retweets	Average Retweet	Sum Favorite	Average Favorite
Q1	1	18	18	9	9
Q2	12	218	18,17	55	4,58
Q3	12	74	6,17	62	5,17
Q4	26	440	16,92	263	10,12
2013	51	750	14,7	389	7,6

Table 24: New York Times Graphics - 2013

Q1: While the last quarter (Q4 2012) rolled out a number of 19 tweets the first quarter of 2013 cannot compete with this amount and only provided one relevant tweet. This one tweet was retweeted 18 times and was liked nine times, which is compared to the last quarter a massive loss in numbers.

Q2: Nevertheless, with twelve relevant tweets and a sum of 218 retweets and 55 favorite clicks, it marks the beginning of an ascending trend of posts with a visual attraction of a data journalistic product.

Q3: The amount tweets in this quarter stay the same as in the last quarter showing twelve relevant posts. The difference is the amount of retweets, which decreased by over 60% down to 74. On the other hand, there is a gain in favorites by around 13% up to 62 star markings.

Q4: 26 tweets mark the highest amount of tweets until here and the posts have been retweeted 440 times, while they were liked 263 times. As a result, Q4 can be called a record quarter up until this point.

2014

In the last year of the period of research, 315 relevant tweets in the four quarters were recorded. These tweets were retweeted 26.370 times and liked 11.760 times, which is a strong increase in numbers and impresses with an amount of quarterly retweets and favorites always above 1.500. With the numbers of Q3 the two categories (retweets and favorite) mark the highest results throughout all the years that were taken into consideration and show a very positive development in the acceptance by the audience. The importance and attraction for the users has also been recognized by the responsible people of *@nytgraphics*, which can be seen by the increase in tweets and high numbers in Q4. The development from 2013 to 2014 is huge and the shares amongst the quarters can be seen here:

	Relevant Tweets	**Sum Retweets**	**Average Retweet**	**Sum Favorite**	**Average Favorite**
Q1	32	2.769	86,53	1.538	48,06
Q2	86	5.460	63,49	2.596	30,19
Q3	74	11.394	153,97	4.276	57,78
Q4	123	6.747	54,85	3.350	27,24
2014	315	26.370	83,71	11.760	37,33

Table 25: New York Times Graphics - 2014

Q1: In relation to the last quarter the increase in tweets (32) is small 23% compared to the upcoming quarters, but the sum of retweets (2.769) and also likes (1.538) shows a big increase that continues over the next quarter.

Q2: The sum of all relevant tweets (86) is almost three times higher than in Q1. Also the amount of retweets went up to 5.460, which means it almost doubled from Q1 to Q2 this year. Similar to this development is the amount of stars that were given to the posts that show an increase from the last quarter to this one by almost 70%, which means in numbers 2.596 likes.

Q3: The positive trend that was drawn throughout the last quarters continues in the third quarter of 2014. Although, only an overall of 74 relevant tweets was provided which is a drop of roughly 14%, the sum of retweets (11.394) and favorites (4.276) show a considerable increase once again.

Q4: In the last quarter that was examined, the amount of tweets rises on an all time high of 123 posts. However, the amount of retweets (6.747) and likes (3.350) decline in comparison to the strong last quarter. In comparison to the other quarters, these numbers still mark the second highest values after Q3.

2.6.3. Summary

After looking at the accounts and the recorded data in detail, the overall development amongst the five different accounts all show a growth in all the categories from mid 2011 until the end of 2014. Overall, the accounts provided 2.963 relevant tweets, of which 45% were posted by *@nytimes* (*New York Times*) and 15% by their specialist (*@nytgraphics*), which leads to 60% (total of 1.710 tweets) of all relevant posts. Second most tweets (20%) were posted by *The Guardian's* specialized account (*@GuardianData*), while *The Guardian* provided 15% of the relevant tweets. The total of both accounts is 1.023 tweets and are more than one-third of all relevant tweets. The least amount of tweets (230 tweets, 8%) provided the German representative *Die Zeit*.

However, along with the amount of tweeted posts the audience's reaction, in form of favorites and retweets was recorded and analyzed. As well as in the case of tweets, the data shows a clear ascending trend in the acceptance, as favorites increased by more than 400 times and retweets by almost 300 times from 2011 to 2014. Once again, the share of retweets amongst the five accounts is dominated by *The New York Times* as the *@nytimes* account, comprises 85% of all retweets and the specialized channel provides 5% of the retweets, which results in 90% of all retweets. Meanwhile, the share of the reposts of the relevant tweets by *The Guardian* is 7% and in combination with its specialized account 9% (2% by *@GuardianData*). The lowest amount of retweets with only 1% is again provided by the account of *Die Zeit*. A similar picture is drawn when looking at the shares of the favorites of the relevant posts. With altogether 92% of all favorite markings *The New York Times* overshadows the others (87 by *@nytimes* and 5% by *@nytgraphics*). *The Guardian* follows up with an overall share of 8% (6% by *@guardian* and 2% by *@GuardianData*), whilst *Die Zeit* has a 1% share of all star-markings. Due to this development, it can be said that the news industry and in particular the pioneers have realized which potential data journalism and its visual products have when sharing it on social networks such as Twitter. Furthermore, this action is rewarded by the audience that shows an increasing interest in these products and form of journalism by liking (favorite) them and further sharing (retweet) them amongst the social groups and community. By that others who do not follow one of the investigated accounts get this data and information as well. The perspectives that open up due to this participation are further described in the conclusion. Additionally, the role that Millennials play and the reasons they might become an important component of data journalism is pointed out.

3. Conclusion

"*Journalists can be the bridge between the providers of data and the consumers - testing, checking and interpreting the data, but also bringing it alive.*" (Rogers, 2013, p.308)

3.1. Conclusion

Since Meyer's pioneering suggestions much has changed in (data) journalism and new components emerged. As described in the introduction, the amount of data exploded and is continuously growing. News media is in a process of transition with new media systems are emerging and the passive consumers become a participatory audience and citizens are able to influence the media, by using technologies to tell their stories in an innovative way. That also affects news media and tighter budgets force newsrooms to restructure in order to (re) act and adapt to the new cultural situation and the fact that the public trust has been lowered over the past years. The upside of these changes and developments are the new tools that are provided and the way digital media offers to create and reach a bigger audience, as more and more people go online and the global network enhances.

Data journalism is currently an interesting field, since it is relatively new and changes constantly due to the emerging possibilities in the technology sector. This chapter summarizes the findings of the thesis and comes to a conclusion what data journalism means for journalists, why it matters and which role Millennials play in connection to data journalism. In order to do that, the research about the characteristics of data journalism in connection to the project (*Dark Horse Vietnam*), the case study and different sources were considered and taken into account. Finally a personal outlook on the development of data journalism in the future is given.

3.1.1. What data journalism means for journalists

The reason why and how data journalism matters can be looked at from two perspectives. One perspectives, shows the advantages as a result of the researched characteristics of data journalism, while the other point is more specific and focuses on the role that Millennials and social networks play in connection to data journalism.

3.1.1.1. Part 1

One positive aspect of data and data journalism addresses the credibility loss, which journalism has been facing over the past years and was already described in the introduction. Due to this development, journalism needs to find a way to tackle this problem and regain credibility and authenticity, to stay competitive, pull readers and win them over as customers. This could be done by engaging more in data, making use of data journalistic stories and picking up the objectives and ideas of Meyer for precision journalism to work journalistically and scientifically with data. Correspondingly to that the journalist publishes sources, such as datasets and background information of the story, the process and the final product. Accordingly, this procedure opens up a spectrum of advantages, but most importantly, through this new transparency, journalism can be considered more trustworthy and credible than the current state of journalism. Furthermore, this change towards scientific or precision journalism offers a way to deal with complex information and find a way to "*(...) communicate essential truth*" (Meyer, 2001, p.4) and stay accountable. So, by passing on information and sources to the audience, data journalism involves and engages the private citizen who is referring to Lippmann (1993, p.3) a "*(...) deaf spectator in the back row (...)*". Moreover, he mentions that the audience or private citizens are aware that they are affected by the rules and regulations, but the public affairs are invisible to them. Additionally, the audience needs to have the ability to understand what is happening directly and grasp the provided information, which Lippmann (1993, p.4) criticizes when pointing out that "*He* [private citizen] *lives in a world which he cannot see, does not understand and is unable to direct*". Because of that, it needs journalism of high quality to increase the citizen's participation in societies and political democratic life (Lippmann, 1965, p.30) and makes them an attentive, active consumer in the first row. This high quality journalism can be provided when fulfilling specific tasks (Kovach & Rosenstiel, 2001, pp.12-13):

1. *Journalism's first obligation is to the truth.*
2. *Its first loyalty is to citizens*
3. *Its essence is a discipline of verification.*
4. *Its practitioners must maintain an independence from those they cover*
5. *It must serve as an independent monitor of power.*
6. *It must provide a forum for public criticism and compromise*
7. *It must strive to make the significant interesting and relevant.*
8. *It must keep the news comprehensive and proportional*
9. *Its practitioners must be allowed to exercise their personal conscience.*

These tasks and requirements are met, when using stories based on data and presenting the findings in a visual appealing and proportional way by for example interactive visualizations. As a matter of fact, data journalism also strengthens the position of the journalist by referring to reliable numbers or as Mirko Lorenz (Gray, 2012, p.4) stated: "*Less guessing, less looking for quotes; instead, a journalist can build a strong position supported by data, and this can affect the role of journalism greatly.*"

Consequently, a more scientific style of reporting could benefit the audience to understand and follow the journalist's way to his or her conclusion. Because, data journalism can attract readers that also want to go beyond the data and analysis and further want to reproduce the conclusions. By investigating and providing background information news media serves and improves its democratic function, which is considering Schudson (1995) a vital function of the news because it is the channel through and with which people think. Moreover, he points out the importance of reporting by the news media in connection to a democratic society in his paper *News, and Democratic Society: Past, Present, and Future* (Schudson, 2008). He made clear that the functions of news are to inform, investigate, analyze, personalize, communicate and mobilize the people, which can be achieved through data journalism and its principles. Moreover, the data journalistic tools empower and support the investigative approach of *watchdog journalism* and help the news media to stay the fourth estate, which plays an essential role in the global democracies, because it documents, exposes and holds people, institutions and governments accountable for failures (Kaplan, 2013) and fulfills the duties of the news media as the fourth estate. A reason Alejandro (2010, p.43) added is:

> "*Journalism has always been about being the vanguards of the community, listening to people, researching facts, giving venue to a road range voices and representing public interest (...).*"

Furthermore, the audience can help to document and communicate the deficiencies through social networks by providing information and data. Butch Ward (2014) emphasizes this movement and stated: ”*(...) We need to invite the community - through the Internet, through partnerships with broadcasting, through forums, to help us with this* [to find solutions not just problems]”. This development improves and progresses digital democracy to create a well-informed, mature and responsible citizen. However, the process to investigate and getting information about backgrounds is benefiting the extension of the open data phenomena and connected to this phenomenon Priya Kumar (2015) gave another raison d'être for data journalism by saying: “*People don't collect data to create charts or graphics; they collect data to answer a question or learn something*”. Therefore, the progression of data journalism is related to the concept of free and open data since it also provides clean datasets that can be explored by the audience. Moreover, the free access and transparency of available data is an important part of *open government* (von Lucke, 2010, p.15).

Still, open data is not enough (Kaplan, 2013) and investigative journalism will remain essential as biased numbers and fake rather than real records will be used and published by the public organs, as they will learn to cover actions from open data. Ben Goldacre (2008) addressed this issue and gives several scientific examples and the ways to make results of trials look positive. As a matter of fact “faking” results and “*publication bia*s” (Goldacre, 2008, p.214) is already practiced in the science society. Due to that, the journalists need to be able to understand data and turn it into a story that serves the public interest without affecting and harming the individual. This thin line became apparent during the revelations of *Wikileaks*. As mentioned before, data such as cables or logs from the *Pentagon* and *U.S. Department of Defense* were given to different news organizations, which had to decide what and to what extend they publish the data. In particular, the protection of individual people by keeping their names secret were key elements. These examples also raise the attention of ethics in this field that will not be further discussed but needs to be mentioned, as it will gain importance on a global level because journalists have to weigh up when they access or scrape data. Also the issue of privacy and security will arise in association with the storage of sensitive data and confidential sources. These two topics will affect the journalist's work and play a major role in the near future. Especially security needs to improve in order to protect the data and enhance and extend *open data* and *open government*.

Consequently, transparency and accuracy should not only apply to governments and public institutions. Moreover, it should be applied to and validated for journalism as well. The reason as Prof. Dr. Klaus Meier stated (2008), is the relationship between journalists and the audience would be increased and become better through transparency, which provides more authenticity and accountability. In the end transparency might be the new objectivity and needs to be pushed forward but the critical thinking will remain essential. Also the accuracy in working with data is significant, because making a mistake during processing data or misinterpreting it, can degrade the premise of precision and data journalism.

Resulting from this transparent work, the journalistic accountability can be restored and the presented characteristics of data journalism benefits, improves and extends the development of citizen journalism but it also benefit data journalism. Moreover, data journalism offers a more reliable way to work transparently and provide a comprehensible and accountable journalism. In particular, as in recent past biased articles and reports were and still are published due to *churnalism* under the cover of journalistic objectivity.

Data project

During the realization process of *Dark Horse Vietnam* different lessons were learned and findings were made. Overall it can be said that with a good story, the appropriate skills and a good team, data journalism is not very different to traditional journalism. It represents the practical evidence of data journalism bringing sense into data and transforming it into a physical product, which can be experienced and helps the citizen to understand abstract incidents such as trade flows. It also pointed out another advantage for journalists to apply data because it assists to discover trends and explore hypotheses. Moreover it helps to move away from a journalism that relies on quotes and leaves the audience unknowing where the truth lies and provides explained data and information, so the reader can explore the truth themselves.

The topic is chosen in comparison to other data stories out there and very basic and simple example. Nevertheless, it particularly provided a chance to test the models of Paul Bradshaw. In conclusion of the use of his model of the *inverted pyramid*, it can be said that it provides a very good guideline and approach to create a data story. In particular one point that needs to be stressed: *context*. It describes the importance to understand the way the mined data was generated, to avoid making use of data resulting "*bad sci-*

ence" (Goldacre, 2008). In the future, this step will become crucial and the skills concerning the scientific gathering of data and the knowledge of statistics are essential to understand the background of the data. The background of this data should also be communicated and described to the audience to keep the journalistic results and conclusion accountable and transparent.

However, one issue during the realization of the article was to communicate the story in the right tone and language. After looking at different examples of data stories on different platforms such as *The Guardian Datablog* or *The Upshot*, the language varies from an informal to a classic journalistic form. In the case of the project *Dark Horse Vietnam* a more informal language was chosen that supports the graphics and helps to understand them and also addresses a younger target group. The idea to use less text was, to let the reader explore the visual products themself and point out only specific data.

Another challenge represented the idea to realize a project based on open data. This type of data made the research easier, but it also became apparent that the sources are still lacking accessible data. Besides the accessibility of the file formats, the definition of the data, which was used by the sources, differed and used different types of quantities and scales or sources of numbers. Nevertheless, after understanding how the data was gathered by the sources, the provided data and definitions can be considered reliable and facilitated the realization of the project.

These findings emphasize once more that open data needs more attention and certain standards have to be set and/ or clear definitions should be provided with the dataset. To give an example, the definition of the terms "unemployed" or "casualty" is unclear, misunderstood and interpreted in the society. In other words and to bring it to a point, there is a lot of data out there but unfortunately also a lot of junk. Therefore, the data needs to be treated with journalistic rigor.

During the realization of the project it could also be recognized that the tools a journalist needs to put data journalism into practice are mostly *open-source software* (OSS), which makes the communication amongst teams easier and more intuitive. Henceforth, the access to this field is open to anyone and the required skills can be learned through online tutorials or *massive open online courses* (MOOC).

An element that adds value and improves the personalization of a topic is interactivity. It supports the journalists to tell cohesive and complex stories and engage and lead the audience, which is an important component because as Lippmann stated it needs "*(...) a personal identification with the stories he* [the reader] *is reading*" (Lippmann, 1965, p.328). As a result he notes that news, which do not make use of this opportunity, will struggle to attract and appeal a wide audience. Through an interactive narrative, the audience can be engaged through its distinctiveness, liveliness, modernity and resulting understanding (Miller & Raisma, 2015). More specifically, interactive applications, utilities and data visualizations help to move beyond static graphics and provide the possibility for the audience to dig deeper into data and spreadsheets by using computers and/ or mobile. Additionally, by letting the user interact with charts or visualization, immediate changes in data and processes can be seen, which leads to a better understanding, personalization and gaining more details. Colin Ware (2004, p.2) called interactive visualizations "*(...) the interface between the two sides*", meaning the visual result of adaptive decision-making and computational power with its massive information resources. Particularly, in connection to the increase in mobile usage, interactive tools represent a supportive and helpful element.

However, in the end it is not about the tool itself but in how far it effectively supports and helps to tell a story. Moreover, the work with data, understanding, scraping and processing needs to be taught to journalists because it needs a lot of care and attention in order to tell the story the right way. As the project shows, data projects can be created quickly over weeks or as more complex news applications prove, it can also take months or even years. Nevertheless, journalists will always have to check the facts and sources because the use of data will not eliminate this task. In fact, the opposite is true because critical thinking, understanding of statistics and data processing is increasingly necessary.

Future skills

During the realization process of the project *Dark Horse Vietnam*, the required skills and also the tools that were used, were rather simple and can be handled by one person. Altogether, the project required basic data-science and visualization skills, which should be the fundamental skills of a data journalist to be able to work as investigative journalist. This includes the construction of databases and leads up to creating the visual and/ or interactive data journalistic products. Moreover, a general understanding of scientific methodologies and statistics should be supplied.

Also *The Poynter Institute of Media Studies* dealt with this subject of the required skill-set of future journalists. The survey was made amongst, professionals, educators, students and independent participants and brought up a list of 37 core skills, which modern journalists need and should be taught, in order to keep up with the change in media and journalism. The report (Finberg & Klinger, 2014) concluded, that multimedia and digital skills, which means the production of graphics, photos and videos as well as the editing of the footage are new essential skills. This also includes basic programming, in order to create small tools or webpages. Another important skill is the need of journalists to understand and have knowledge of the media landscape and business, including marketing aspects such as managing communities, crowdsourcing information and interaction with the audience. Furthermore to the skills, the report points out the significance of specific skills for the news gathering process. Especially, the ability to collect, analyze and synthesize large amounts of data to create compelling stories and graphics that will gain importance for the future (data) journalist.

Along with the investigated skills by Finberg and Klinger (2014), the journalists need to think critical and question facts, numbers and show how the conclusion was reached. In particular, data analysis is prone to bias (Schrager, 2014) and could draw a wrong conclusion if done wrong. Conclusively, the report shows that educators and universities already realized the urge to teach these skills to future journalists and already introduced data journalism courses, which put its emphasis on the mentioned skillset.

3.1.1.2. Part 2

After putting a focus on journalists, the following is concluding why data journalism matters to the audience and how they and the journalists can benefit from data journalism. The participatory culture becomes even more apparent when thinking of today's life and how it circles more and more around mobile applications, since almost everybody can record videos, take photos and share them with the whole world on social networks. Due to that, the production but also consumption of stories has changed and stories are consumed third, fourth or fifth hand (Alejandro, 2010, p.9). Accordingly, the second part focuses on Twitter as a representative of a social network and the Millennials as heavy user of them (Duggan et al., 2015).

The role of the Millennials

The Millennials represents the generation that soon will take over key positions in economy, politics and media. Due to that they are the point of interest and part of many studies. However, there are several reasons why they will play an important role in data journalism. As the described characteristics of the Generation Y show, they question and scrutinize everything around them and do this primarily online. This information and answer gap of the Millennials can be filled with data journalism that provides stories and sources to educate and pass on news and knowledge. Moreover, everybody can interact with the journalist, can research the data themselves and recreate the data. These mechanisms within media presentation are in favor of the Millennials, because they are considered adventurers and want to explore, know more about backgrounds, which can be achieved by providing and presenting appealing, trustworthy information or data. Additionally, they have questions that might not be completely answered by the data story and by giving them access to the sources; they can explore the world beyond the story and even find a new story in the data. Due to that benefit could be that a data story is more appealing to the young audience because it does not tell them how and what they should think by writing biased stories, *churnalism* that is based on press releases of companies or political parties.

Overall, the connectivity is another key element of the Millennials, which means they want to be connected to a story and perhaps the author, in order to ask questions or communicate their findings through the interactive surface of Twitter or other social media platforms and channels. This brings up the way Generation Y forms their opinion, because other than their precedent generations, they are influenced by different types of people and to a lesser extend on “experts” like financial advisors or doctors. Moreover, they rely on suggestions by friends and are connected to brands (Barton, Koslow & Beauchamp, 2014), which makes Twitter and other social networks fertile soil.

To conclude, the Generation Y will play an important role in the field of data journalism, motivated by their characteristics, knowledge and connectivity. By making use of these attributes, information and news are spread easier. Additionally, crowd sourcing social networks can provide data and information by mining them. The use of this swarm intelligence of the civil society could back up government data and misconduct can be uncovered. Furthermore, their knowledge of technology and enthusiasm to explore new fields can help to provide knowledge, information and sources by for example giving access to archives, which leads to an increase in citizen participation and enables citizen journalism. Due to the gaining participation of the user on social networks, the mentioned watchdog journalism can also be applied by giving feedback, commenting, providing ideas and data to find a new story and monitoring institutions through social media.

Nevertheless, the technological knowledge and appeal to technology does not apply to Generation Y journalists as a study by Daniel Hobbs (2008) showed. He points out that Millennial journalists still rely on traditional news skills, which emphasizes once again the urge to teach skills related to new media and adapt to this "new" form of journalism. Although, Hobbs drew this picture I am optimistic that the change is already happening and journalists from this cohort know and are keen to work with the new tools and push data journalism forward.

As mentioned before, news media has already realized the attractiveness of data journalism, which was also described by Simon Rogers (Gray, 2012, p.139). Still, news media needs to make more use of visuals in connection to data journalism on Twitter to capture the interest of the "non-data-nerds" who can also give input or even produce a story and take their part in the democratization process. So, it is important to learn how to deal with the Generation Y but it is even more important to keep the upcoming Generation Z in mind because they use and cross-over different social networks, like *Snapchat*, *Facebook* and *Twitter* (Lenhart, 2015). In addition, the so-called "*iGeneration*" (Bulik, 2011) has an even higher affinity to technology than the Millennials. Furthermore, the trend of relying less on traditional news is going to continue (Pew Research Center, 2015) and the new possibilities provided by the Internet and its tools need to be used to reach the Millennials and the following generation. Correspondingly, there is an urge for news media to collaborate with Millennials in order to lay and consolidate the foundations by developing data journalism further and attract the Millennials now and the arising *iGeneration* in the near future.

Case Study

Referring to John Snow (Alejandro, 2010, p.33) Twitter is an impacting contribution to journalism and the promotion of the content. Interactivity, Twitter and social media in general, offer a democratization of what news media does and the audience wants to know and further good journalism has always been about networking where journalists have to listen, research and converse (Beckett, 2008, p.46). Therefore, social networks are a suitable platform for journalists to use, because the audience is personalized, fragmented and diversified and can it be used as a platform to inform but also to gather information.

The case study is by no means perfect since the focus was only put on three different, pioneering news media companies. Nevertheless, it served its purpose as an attempt to get an impression of the development of the use of data visualizations on Twitter among news media and the acceptance of visual data journalistic products by the audience. Furthermore, since Millennials are heavy users of this platform it also enables to link it to this group of current and future interest (Mitchell & Guskin, 2013). Before, getting into detail about this cohort, the focus is put on the producers, the three observed news companies.

The general outcome of the case study is that Twitter represents a suitable platform for data journalism. Although, it competes with platforms such as *Instagram*, *Snapcha*t and *Facebook*, which are more popular amongst Millennials (Richter, 2014) and have bigger user groups, Twitter's immediacy, conciseness and way to consume news are its advantage. It can be said that data visualizations have the ability to create more appealing and interesting stories, which are more attractive to the consumer and might lead to an increase of reader because of the fact that some people react more to a visual story than a written one. Overall, the ascending numbers in related tweets throughout the accounts and over the years, show a rising interest and awareness level among the news industry to adapt data journalism and publish data visualizations. This growth cannot be seen to the same extend on every account. Especially, the German *Die Zeit* provided lower numbers than the other two "main" accounts of *The New York Times* and *The Guardian*. One of the reasons is the fact that the German tweets only reach a limited group of users and another one is the amount of active Twitter users in Germany, which is very low compared to its population and other countries (Statista, 2015). The low amount of relevant tweets further implies that the German representative seems to have a lack in resources to work in this field and/ or is not putting so much emphasis on working with Twitter. This assumption is also supported by the number of all of the posted tweets of *@zeitonline*. Nevertheless, the reasons and assumptions are all hypothetical and with certainty it can only be said that *Die Zeit* should start to engage more in social networks and explore the field of data journalism, make more use of the new tools and options to appeal the audience and transfer knowledge and information.

When looking at the four Anglophone accounts the increase in tweets over the years is huge. *The New York Times* and *The Guardian*, both realized which potential and importance data journalism and its visual products have. Due to that, they founded specialized platforms (*The Upshot* and *Datablog*) and offered their specialists a Twitter channel with growing success. Both parties know how to promote their data stories and are omnipresent on social networks and accepted by the audience. In particular the numbers of *The New York Times* let assume that they have the capacity to deal with the new form of journalism and push it further forward by experimenting and trying new ideas. Other than *The New York Times*, is *The Guardian* because they work with fewer resources but still tweeted far more relevant posts than *Die Zeit* and also run a specialized website and account. However, also *The Guardian* with Simon Rogers is able to tackle and support the development of data journalism. The two different capabilities are picked up in the *Advice* and support the mentioned assumptions.

After pointing out that news media has a certain interest to work with data journalism and publish the visual data products on Twitter, the focus is now put on the feedback of the audience.

As well as the development of the tweets, the feedback developed positive and the numbers of favorites and retweets are rising over time, which implies a growing acceptance and interest by the users in visual results of data stories. Even though one reason for this ascend is very likely the increase in Twitter users since 2011 (Statista, 2015), which is three times more in 2014 (Q4) than in the third quarter of 2011. Nevertheless, the smallest growth of favorites and retweets, which was provided by *@GuardianData*, in the same period of time shows 12,5 times more favorites and over 22 times more retweets. So, the increase in interest is not only caused by the growth of Twitter users and also caught the interest of “old” Twitter users.

As a result of the collected data and trends, it can be concluded that the audience of Twitter are attracted by data visualizations and think the information is worth to share in the *twittersphere*. Due to this interest, Twitter provides a good platform to experiment with data journalism and corporate with the user, which can also enable and support the development of citizen journalism. Especially, Millennials empower citizen journalism by participating and engaging with news media through different platforms. As seen, the participation and feedback of the audience was a part of the case study, by recording favorites and retweets to show the interaction between the audience and the news media. Coupled with this interaction and dispersion of knowledge and information is the term citizen journalism. The term goes beyond sharing articles and Jay Rosen, Professor of Journalism at NYU stated that citizen journalism applies, "[w]*hen the people formerly known as the audience employ the press tools they have in their possession to inform one another*" (Rosen, 2008). Later, the author Daniel Bennett adjusted this definition by emphasizing the publication and adding "*many*" and pointing out the importance of a "*newsworthy event*" with this result: "*When the people formerly known as the audience employ the press tools they have in their possession to inform many others of a newsworthy event, that's citizen journalism*" (Bennett, 2008).

To sum it up, citizen journalism involves news and commentary at a larger scale because the audience is making use of *wikis*, *blogs* or social networks as a tool to produce content. Subsequently, this can be used by the news media to contribute and work in "collaborative citizen journalism". Nowadays, news institutions adapted to this new situation by merging and combining the print and online newsrooms and incorporate more citizen journalism or reader generated content. But not all citizen media is citizen journalism, especially when looking at the statistic (Smith & Wollan, 2011, pp.XI-XV), which shows the average content of tweets it can be said: Most is not!

So, it can be said that citizen journalism, as it is known today is a result of the democratization of media and the digital age. Moreover, "*(...) every citizen is not just a potential source but also a potential reporter*" (Auciello, 2013). This new form also brought up citizen media organizations like *ProPublica* that produces investigative journalism and makes: "*Journalism in the Public Interest*" (ProPublica, 2015). Together with the increased usage of social media and networks, this shift from the traditional media outlet towards people-driven news is supported. Moreover, it gives the citizens tools of judgement and a possibility to engage and affect public life, because as Stephen Jay Gould stated: "*When people learn no tools of judgement and merely follow their hopes, the seeds of political manipulation are sown*" (Goldacre, 2008, p.253). However, since the audience decides which news they are attracted to and to what extend they want to engage, todays information are spread bottom-up instead of top-down, as in the past (Bruns, Highfield & Lind, 2012, p.26). Henceforth, another trend developed - networked journalism. It opens new sources by mining the public knowledge, sharing and connecting continuously as the journalist reports. By that, the role moved away from being a gatekeeper to becoming a connector (Beckett, 2008, p.56). Regardless, both groups serve the same core practice, they are gate watching (Bruns, 2005), which means they monitor newsworthy organizations, gather and compile data and information, process and publish them so news stories and comments can be produced that rely on these information and data (Bruns, 2012, p.26-27).

As technological innovations rapidly emerge and happen, the news industry needs to keep up with them. In the beginning they might disrupt the news process but will change the news industry and bring up hybrid forms of journalism such as data journalism. In the end, it is about the presentation of news, which needs to appeal to the audience. This can be achieved by combining the principles of data journalism and the tools of visual and computational journalism. The presentation is an essential aspect in order to reach the Generation Y and interact with them, as the news media has to compete with platforms such as *BuzzFeed*. When implementing data journalism in newsrooms successfully, news companies are able to brand their media company and connect to the audience in a unique way, which creates an exclusive way to give and get data, information and knowledge. Nonetheless, up until now data journalism has not reached its full potential, but examples like the *Datablog* show that there is a niche in journalism, which is accepted and appreciated by the audience. As a result, the field of data journalism is going to influence the journalists and newsrooms.

3.2. Advice

After showing the advantages in connection to recent developments and what the journalists need to adapt to, this part takes a practical viewpoint. In the following, ways and approaches to implement data journalism in the newsrooms are presented and explained. The reason is that newsrooms need to adjust and to tackle the changes in media and practice data journalism.

3.2.1. Integration Newsroom

Newsrooms need to include data journalists in their newsrooms or make use of external data journalists in the near future to stay competitive and innovative. In order to achieve that, they need to be transformed into collaborative newsrooms, which collects, analysis, shares and uses data to create stories and insights (Zanchelli & Crucianelli, 2012). That means the newsroom needs to become a multimedia newsroom, which includes the different elements that it takes to create a successful data journalism task force. These teams should be supported or consist of Millennials because they can add value by addressing new topics and bring in a diversity of "technological" skills (The Guardian, 2014). Michael Zanchelli with Sandra Crucianelli (2012) of the Knight International Journalism Fellowship dealt with the idea of a suitable newsroom and pointed out four essential aspects that play a significant role to successfully integrate data journalism in newsrooms.

1. Proximity to news desk
   This gives the data journalism team critical access to editors and reporters for feedback on data projects. Furthermore, collaboration between the data team and news desk can be generated and improved.

2. Motivate reporter and developer to create data stories
   Due to different specializations both need to cooperate to bring up story ideas because they are likely to find unique and significant angles for the data story.

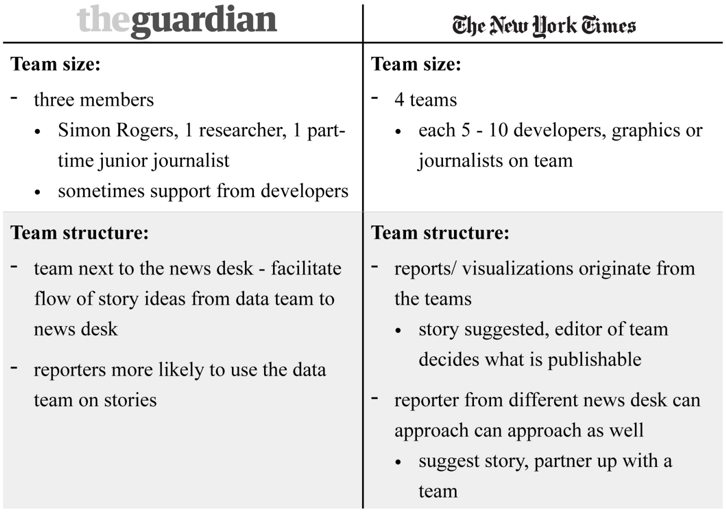
3. Recruit people who bridge the skills gap
   People with skills in journalism, data mining or coding are valuable assets to the team because they provide new possibilities and approaches.

4. Produce stories that engage the audience
   Data stories about topics that affect the lives of the consumer create impact and web traffic, which emphasizes the need to work with data journalism teams.

Meanwhile, the size of the teams and their location differs as Zanchelli shows in his article as well as summarized in fig. 20.

As mentioned before the capacities and resources of two of the relevant news media organizations for the case study, *The Guardian* and *The New York Times* differ and so they also have different approaches to tackle data stories. Even though, journalists are accompanied by specialists, the journalist still needs a general understanding of the research methodology, if they want to succeed in data journalism. Certainly, the two examples given above are big newsrooms, which are also pioneering in the field of data journalism.

Fig. 20 - Comparison newsrooms



theguardian	The New York Times
Team size: - three members • Simon Rogers, 1 researcher, 1 part-time junior journalist • sometimes support from developers	**Team size:** - 4 teams • each 5 - 10 developers, graphics or journalists on team
Team structure: - team next to the news desk - facilitate flow of story ideas from data team to news desk - reporters more likely to use the data team on stories	**Team structure:** - reports/ visualizations originate from the teams • story suggested, editor of team decides what is publishable - reporter from different news desk can approach can approach as well • suggest story, partner up with a team

However, also small newsrooms can make use of data journalism. They will not be able to provide data stories on a high frequency, like the two newsrooms mentioned above because the teams would be smaller and the data-collection process is time-consuming and difficult. As a result of working in smaller teams, the group needs to take on multiple roles within the process, which results in slower processing of the data story. Due to that, the team needs a different structure than the examples provided above. In fact, the journalist needs to be extremely selective and gather a large amount of data to select from, in order to put specific events in a short story (Meyer, 2002, p.4).

Nevertheless, a smaller team is able to produce data stories as well, when recruiting the right people that provide and combine the skills mentioned. Also the gadgets are already available on the Internet, which provides powerful tools to scrape, process and visualize data. As a result, even smaller newsrooms with limited resources can tackle the challenge data journalism.

Computational Journalism

A helpful tool for small newsrooms or local news media is computational journalism because it can increase the capacity of newsrooms by automating tasks. This form of journalism is tightly connected to data journalism and it provides research possibilities and the ability to automated, machine written articles. This was the case in March 2014, when the "*Quakebot*" generated the first news report (BBC News, 2014) for the Los Angeles Times. Later that year, at the *Computation + Journalism Symposium* the paper *Journalist versus news consumer: The perceived credibility of machine written news* (Krahmer & Van der Kaa, 2014) was presented. The paper deals with the perception of robot-generated news and shows that there are no differences in perception regarding credibility and readability of robot-written news articles. Hence, automated news about for example earthquakes, traffic, sports or police blotters can be used in local newspapers to pull readers to their sites with a minimum effort of programming the algorithm.

Additionally, computational journalism provides a wide range of possibilities to do data research. Newsbots could be generated as an alert system for beat reporters, as the concept of "*RumorLens: A System for Analyzing the Impact of Rumors and Corrections in Social Media*" (Resnick et al., 2014) shows. *RumorLens* is a tool that was made to help journalists to identify new rumors on Twitter. There are many more ways to implement computational journalism to support and make the work of a data journalist easier, like applications that help visualize the results of data stories such as *Many Eyes*.

Moreover, news applications that will provide a journalistic service, so people can make better-informed decisions will gain importance in the newsrooms. These applications represent a new and important form of storytelling or narrative, which is accessible across all devices and operating systems. The term refers to interactive features accessible through online applications. Accordingly, this feature offers a suitable way to deliver bigger data stories and information to mobile devices.

In summary, journalism has to adapt to computation due to the quantitative turn society is taking and data journalism is a beginning of this "*(...) hybridization of journalism and the computing and data sciences*" (Krisch, 2014). In the future, data journalism and computational journalism are very likely to merge because as data journalism is growing, the field of expertise of computational journalism is going to be adopted and transferred, to enlarge the computational work of data journalism.

3.2.2. Outlook and Perspectives

It is hard to say how journalism and especially data journalism will look in ten years or more but it can be speculated how it could develop in the near future. Some ideas and developments were already mentioned like the potential influence of computational journalism. However, the following is describing further ideas and perspectives based on the findings of this thesis.

Overall, data journalism will not be the messiah to save the journalistic world but it certainly will change it and bring new possibilities and tools to the journalists to make their job easier and less vulnerable by working transparent. Philip Meyer's idea of precision journalism is based on social science as a tool for investigative reporting and in order to do so, the journalists need to understand the methodology. The previously presented components and elements will help to improve data journalism and further to engage the audience in the future. Moreover, it is important to educate and make data more accessible and understandable to the mass (readers), especially in times where governments talk about data security and transparent citizens, so public is not frightened of the unknown and more aware and competent to judge and discuss such topics and its backgrounds. In particular, the open data movement needs to be pushed forward in the near future. In connection to that, publishers should digitize their archives (Ciobanu, 2015) to make historical data and footage available, in which new stories can be found in old data or can be connected to new stories and assist to put them into perspective. In association to the expansion of available open data, data journalism will gain more importance in the newsroom and will certainly become an important component because it currently is the most "objective" and factual resource that assists journalism to regain trust and credibility for the sector and media if done the right way. Therefore, the journalists need to and will adapt to this situation.

The main abilities always were writing, interviewing and checking facts their work and sources. By now, photography, videography and social media are already integrated in the toolkits of most journalists. Especially, as more people move online with mobile devices and the number tablets, smartphones and phablets is growing, these tools and particularly interactive news applications using multimedia will help to design and tell stories in a new way, by relating the narrative to the user. Also, as reporting needs to be faster, due to networks like Twitter the automated processes such as drone journalism or robot/ computational journalism will grow in use. For instance, less profitable beats such as traffic, could be covered by computers or drones that gather data and keep users informed almost in real-time.

Nevertheless, data journalism will also face rough times, created by bad data journalism, as more big data will be produced by various sources to discover and predict the future of almost every field (Lohr, 2015). In reference to the *problems of data journalism*, data can be erroneous or biased due to faulty, deficient experiments or the motivation of company to have a specific outcome of a study. In the following the bad data will be used without critical thinking and mislead and misinform the audience, because journalists do not apply their traditional skills and simply believe the numbers without questioning them. Of course, this can also happen with "good" data and critical thinking if the journalists do not have the toolkit to deal with data analysis or data presentation. As a result, journalists will be trained in this field so they can enrich and support their stories with statistics, social science methods and data visualizations. Furthermore, modern media organizations could become institutions that provide credible and trustworthy data because "good" data will become an asset. Especially news agencies are in advantage since they already host and archived large amounts of data.

Of course, there are also critical opinions on data journalism. Some see it as a trend and hype such as David Leonhardt, managing editor at *The New York Times*, who described it as a "*fashionable*" phrase. He thinks data and statistics *"(...) are both the most overrated and underrated kind of information*" (Leonhardt, 2015). Additionally, he points out that in reality big data and its analysis can be wrong because statistics can be used to achieve a specific aim, but takes into account that "*'Statistics' and 'data' are really just a plural form of 'fact'*" (Leonhardt, 2015) and seeks to be "fact journalism". Leonhardt's criticism correlates and takes up the problem of bad data and dealing correctly with data in general.

Considering this criticism and point of view, data journalism should not only rely on scientific methods and needs be complemented with traditional skills and wisdom journalism (Stephens, 2014) to create successful and compelling stories. The term wisdom journalism, describes a form of reporting that is "*(...) exclusive, enterprising, investigative*" (Stephens, 2014, p.XXVI), which further "*(...) emphasizes informed, interpretive, explanatory, even opinionated takes on current events*" (Ibid.). This wisdom can be gained by collaborating with academics and institutions such as libraries that provide knowledge, data and also the information of witnesses. Therefore, a combination of these characteristics with the tools and principles of data journalism could pair up and go beyond news, reporting what has happened and provide background data and facts in a certain quality, because this "*quality journalism*" provides the information that is needed in order to be an engaged citizen (Stephens, 2014, p.1).

3.2.3. Suggestions future research

The field of data journalism offers a wide range to do further research. In connection to this thesis it is interesting to look further at the change of the newsrooms but also the journalist and its skills. However, in my opinion the most appealing field and suggestion for a future research is the cross-border-collaboration of journalists and data sources such as *WikiLeaks*. Due to that, one approach in the future is to research in a comprehensive crossover of political studies and journalism in order to find ways to simplify the communication regarding the exchange of data and information. I am particularly interested in the consequences of the democratization of data for governments and related civic journalism, the revival of the watchdog journalism and the media as the fourth estate within a country. Another research, which I want to emphasize, is related to the field of marketing and journalism. The goal in this case is to think of possible ways to finance and support data journalism, to keep up the information flow and pressure on public institutions to publish data and democratize the date.

4. References

Due to data journalism's "young" age and its recent appearance, it is fundamental make use of a range of books, papers, interviews and opinions of people from the scenery. In the following, the sources, further information and a collection of data from the case study can be found.

4.1. Bibliography

Alejandro, J. (2010). *Journalism in the Age of Social Media*. [Online]. Available at <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Journalism%20in%20the%20Age%20of%20Social%20Media.pdf>. [Accessed on 24/05/2015].

Alexa. (2015). *Alexa Top 500 Global Sites*. [Online]. Available at <http://www.alexa.com/topsites>. [Accessed on 24/03/2015].

Auciello, J. (2013). *The stronger citizen reporter*. [Online]. Available at <http://www.niemanlab.org/2013/12/the-stronger-citizen-reporter/>. [Accessed on 22/05/2015].

Bardoel, J. and Deuze, M. (2001). *Network Journalism: Converging Competences of Media Professionals and Professionalism*. [Online]. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.8231&rep=rep1&type=pdf>. [Accessed on 14/05/2015].

Barton, C., Fromm, J. and Egan C. (2012). *The Millennial Consumer. Debunking Stereotypes*. [Online]. Available at <http://www.bcg.com/documents/file103894.pdf>. [Accessed on 20/03/2015].

Barton, C., Koslow, L. and Beauchamp, C. (2014). *How Millennials Are Changing the Face of Marketing Forever*. [Online]. Available at <https://www.bcgperspectives.com/content/articles/marketing_center_consumer_customer_insight_how_millennials_changing_marketing_forever/>. [Accessed on 15/05/2015].

BBC Academy. (2015). *BBC Academy - Journalism - Visual journalism: Motion graphics*. [Online]. Available at <http://www.bbc.co.uk/academy/journalism/article/art20141021160021843>. [Accessed on 29/05/2015].

BBC News. (2014). *Robot writes LA Times earthquake breaking news article*. [Online]. Available at <http://www.bbc.com/news/technology-26614051>. [Accessed on 05/05/2015].

Beckett, C. (2008). *Supermedia: Saving journalism so it can save the world*. Malden, Mass., Wiley.

Belmonte, N. (2013). *Visualised: the world responds to Typhoon Haiyan on Twitter*. [Online]. Available at <http://twitter.github.io/interactive/philippines/> [Accessed on 06/06/2015].

Bennett, D. (2008). *Mediating Conflict: The definition of citizen journalism considered*. [Online]. Available at <http://www.dsbennett.co.uk/2008/07/ignore-that-last-post-defintion-of.html>. [Accessed on 22/05/2015].

Bergh, J. and Behrer, M. (2011). *How cool brands stay hot branding to Generation Y*. London, Kogan Page.

Biocca, F. A. (1988). Opposing conceptions of the audience: The active and passive hemispheres of mass communication theory. *In*: Anderson, J.A. (Ed.). *Communication Yearbook 11*. Thousand Oaks, Calif., Sage Publications, pp. 51-80.

Boston Consulting Group. (2013). *How Millennials Are Changing the Face of Marketing Forever*. [Online]. Available at <https://www.bcgperspectives.com/content/articles/marketing_center_consumer_customer_insight_how_millennials_changing_marketing_forever/?chapter=3>. [Accessed on 18/05/15].

Bradshaw, P. (2011). *6 ways of communicating data journalism (The inverted pyramid of data journalism part 2)*. [Online]. Available at <http://onlinejournalismblog.com/2011/07/13/the-inverted-pyramid-of-data-journalism-part-2-6-ways-of-communicating-data-journalism/>. [Accessed on 16/05/2015].

Bradshaw, P. (2011). *The inverted pyramid of data journalism*. [Online]. Available at <http://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/>. [Accessed on 16/05/2015].

Bruns, A. (2005). *Gatewatching: Collaborative Online News Production*. New York, N.Y., Peter Lang.

Bruns, A., Highfield, T. and Lind, R. A. (2012). *Blogs, Twitter, and breaking news: The produsage of citizen journalism*. *In*: Lind, R. A. (Ed.). *Produsing Theory in a Digital World: The Intersection of Audiences and Production in Contemporary Theory 80*, New York, N.Y., Peter Lang, pp. 15-32.

Bulik, B. S. (2011). *The iGeneration: There's a Market for That - and It's a Big, Influential One, Too*. [Online]. Available at <http://adage.com/article/news/igen-influential-peers-household-buying-decisions/230427/>. [Accessed on 23/04/2015].

Bullock, A. (2012). *Kaiser Fung on data analysis & information visualization*. [Online]. Available at <http://blogs.sas.com/content/jmp/2012/10/09/kaiser-fung-on-data-analysis-information-visualization/>. [Accessed on 30/03/2015].

Cairo, A. (2013). *The Functional Art: An introduction to information graphics and visualization*. (Kindle ed.). Berkeley, Calif., New Riders.

Card, S., Mackinlay, J. and Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco, Calif., Morgan Kaufmann.

Chiasson, T. and Gregory, D. (2014). *Data + Design: A Simple Introduction to Preparing and Visualizing Information*. Columbia, Mo., RJI.

Ciobanu, M. (2015). *Why publishers should 'bite the bullet' and digitise archives*. [Online]. Available at <https://www.journalism.co.uk/news/why-publishers-should-bite-the-bullet-and-digitise-their-archives-/s2/a565254/>. [Accessed on 02/05/2015].

Deacon, T. (1998). *The Symbolic Species: The Co-evolution of Language and the Brain*. New York, W.W. Norton.

Dorsey, J. (2011). *search+photos*. Twitter Blogs. [Online]. Available at <https://blog.twitter.com/2011/searchphotos>. [Accessed on 23/04/2015].].

Duggan, M. and Brenner, J. (2013). *Social Networking Site Users*. [Online]. Available at <http://www.pewinternet.org/2013/02/14/social-networking-site-users/>. [Accessed on 26/05/2015].

Duggan, M. and Brenner, J. (2013). *The Demographics of Social Media Users - 2012*. [Online]. Available at <http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/>. [Accessed on 26/05/2015].

Duggan, M. et al. (2015). *Social Media Update 2014*. [Online]. Available at <http://www.pewinternet.org/2015/01/09/social-media-update-2014/>. [Accessed on 26/05/2015].

Egawhary, E. and O'Murchu, C. (2012). *Data Journalism (CAR)*. The Centre For Investigative Journalism. [Online]. Available at <http://issuu.com/tcij/docs/data_journalism_book/1?e=2989993/2615602>. [Accessed on 19/04/2015].

Erny, S. (2013). *Zahl der Bahnunfälle geht seit Jahren zurück*. [Online]. Available at <http://www.srf.ch/news/schweiz/zahl-der-bahnunfaelle-geht-seit-jahren-zurueck>. [Accessed on 27/05/2015].

European Commission. (2003). *Directive 2003/98/EC of the European Parliament and of the Council*. [Online]. Available at <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF>. [Accessed on 16/05/2015].

European Commission. (2015). *European legislation on reuse of public sector information.* [Online]. Available at <http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information>. [Accessed on 07/05/2015].

Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten*. Oakland, Calif., Analytics Press.

Finberg, H. I. and Klinger, L. (2014). *Core Skills for the Future of Journalism*. [Online]. Available at <http://www.newsu.org/course_files/CoreSkills_FutureofJournalism2014v2.pdf>. [Accessed on 23/06/2015]

Fung, K. (2009). *Worthy of the Times*. [Online]. Available at <http://junkcharts.typepad.com/junk_charts/2009/11/worthy-of-the-times.html>. [Accessed on 27/05/2015].

Fung, K. (2010). *Five Years of Chart Reading*. [Online]. Available at <http://junkcharts.typepad.com/junk_charts/images/JunkChartsNYUTalkApr2010.pdf>. [Accessed 27/05/2015].

Fung, K. (2010). *Junk Charts talk*. [Online]. Available at <http://junkcharts.typepad.com/junk_charts/2010/05/junk-charts-talk.html>. [Accessed on 27/05/2015]

Gelman, A. (2009). *Senators and health care; also a discussion of pretty statistical graphics - Statistical Modeling, Causal Inference, and Social Science*. [Online]. Available at <http://andrewgelman.com/2009/11/19/senators_and_he/>. [Accessed on 12/03/2015].

GfK Verein. (2014). *Trust in Professions 2014*. [Online]. Available at <http://www.gfk.com/Documents/Press-Releases/2014/GfK_Trust%20in%20Professions_e.pdf>. [Accessed on 30/05/2015].

Glaser, B. and Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for qualitative research*. Chicago, Ill., Aldine Pub.

Goldacre, B. (2009). *Bad Science*. London, Fourth Estate.

Gray, J. (2012). *The Data Journalism Handbook: How Journalists Can Use Data to Improve News*. Sebastopol, Calif., O'Reilly Media.

Grittmann, E. (2001). Fotojournalismus und Ikonographie. Zur Inhaltsanalyse von Pressefotos. *In*: Wirth, W. (Ed.). *Inhaltsanalyse: Perspektiven, Probleme, Potentiale*. Köln, Herbert von Halem , pp. 262-279.

Grittmann, E. and Ammann, I. (2009). Die Methode der quantitativen Bildtypenanalyse. Zur Routinisierung der Bildberichterstattung am Beispiel von 9/11 in der journalistischen Erinnerungskultur. *In*: Petersen, T. and Schwender, C. (Ed.). *Visuelle Stereotype*. Köln, Herbert von Halem, pp. 141-158.

Hart, J. C. (2015). *Introduction*. (Video file). The Computer and the Human. On coursera. Data Visualization. 2015. University of Illinois at Urbana-Champaign. Available at <http://d396qusza40orc.cloudfront.net/datavisualization/recoded_videos%2F01-11-intro%20take1.mpg.f928d8d02e6011e5baea93ce5eea43d5.webm>. [Accessed on 12/06/2015].

Hart, J. C. (2015). *Overview of Visualization.* (Video file). The Computer and the Human. On coursera. Data Visualization. 2015. University of Illinois at Urbana-Champaign. Available at <http://d396qusza40orc.cloudfront.net/datavisualization/recoded_ videos%2F01-13-overview.mpg.aaf4bf202e6111e5b3bcd94ead694dbf.webm>. [Accessed on 12/06/2015].

Hartmann, F. and Bauer, E. (2006). *Bildersprache: Otto Neurath, Visualisierungen*. Wien, WUV.

Harvard University. (2015). *Survey of Young Americans' Attitudes toward Politics and Public Service*. [Online]. Available at <http://www.iop.harvard.edu/sites/default/files_ new/IOPSpring15%20PollTopline.pdf>. [Accessed on 03/06/2015].

Hernandez, R. (2013). *Those required courses in journalism school are there for a reason*. [Online]. Available at <http://www.niemanlab.org/2013/10/robert-hernandez-those-required-courses-in-journalism-school-are-there-for-a-reason/>. [Accessed on 22/05/2015].

Hobbs, D. (2008). *Gen Y Journalists*. [Online]. Available at <http://ejournalist.com.au/ v8n2/Hobbs.pdf>. [Accessed on 16 Aug. 2015].

Holmes, D. (2013). *Flowchart: Should journalists learn to code?*. [Online]. Available at <https://pando.com/2013/10/23/flowchart-should-journalists-learn-to-code/>. [Accessed on 18/05/2015].

Holyoak, K. (2012). *The Oxford Handbook of Thinking and Reasoning*. Oxford, Oxford University Press.

Hurrelmann, K. and Albrecht, E. (2014). *Die heimlichen Revolutionäre. Wie die Generation Y unsere Welt verändert*. Weinheim/ Basel, Beltz Verlag.

Kaplan, D. E. (2013). *Global Investigative Journalism: Strategies for Support*. Center for International Media Assistance. [Online]. Available at <http://www.cima.ned.org/ wp-content/uploads/2015/02/CIMA-Investigative%20Journalism%20-%20Dave%20 Kaplan.pdf>. [Accessed on 12/06/2015].

Kaplan, D. E. (2013). *Why Open Data Isn't Enough.* [Online]. Available at <http://gijn. org/2013/04/02/why-open-data-isnt-enough/>. [Accessed 02/06/2015].

Keeter, S. and Taylor, P. (2009). *The Millennials*. [Online]. Available at <http://www. pewresearch.org/2009/12/10/the-millennials/>. [Accessed on 18/05/2015].

Khatchadourian, R. (2010). *No Secrets*. [Online]. Available at <http://www.newyorker.com/magazine/2010/06/07/no-secrets?currentPage=4#ixzz0pi3DCk1g>. [Accessed on 09/05/2015].

Kirk, A. (2012). *Article: The 8 hats of data visualisation design - Visualising Data*. [Online]. Available at <http://www.visualisingdata.com/2012/06/article-the-8-hats-of-data-visualisation-design/>. [Accessed on 12/05/2015].

Kirk, A. (2012). *The 8 Hats of Data Visualisation*. [Online]. Available at <http://www.slideshare.net/visualisingdata/the-8-hats-of-data-visualisation>. [Accessed on 03/05/2015].

Kovach, B. and Rosenstiel, T. (2001). *The elements of journalism: What newspeople should know and the public should expect*. New York, N.Y., Three Rivers Press.

Krahmer, E and Van der Kaa, H. (2014). *Journalist versus news consumer: The perceived credibility of machine written news*. [Online]. Available at <http://compute-cuj.org/cj-2014/cj2014_session4_paper2.pdf>. [Accessed on 13/06/2015]

Krämer, W. (1999). *Wie schreibe ich eine Seminar- oder Examensarbeit?*. Frankfurt a.M., Campus.

Krisch, M. (2014). *2014 C+J Symposium*. [Online]. Available at <http://computation-and-journalism.brown.columbia.edu/>. [Accessed on 09/02/2015].

Kumar, P. (2015). *Using Data Storytelling to Engage Audiences*. [Online]. Available at <http://www.pbs.org/idealab/2015/03/using-data-storytelling-to-engage-audiences/>. [Accessed on 27/05/2015].

Kwak, H. et al. (2010). *What is Twitter, a Social Network or a News Media?*. [Online]. Available at <http://an.kaist.ac.kr/~haewoon/papers/2010-www-twitter.pdf>. [Accessed on 14/05/2015].

Lankow, J. (2012). *Infographics: The power of visual storytelling*. Hoboken, N.J., Wiley. (Kindle ed.).

Larkin, M. (1999). *Earliest Egyptian Glyphs - Archaeology Magazine Archive*. [Online]. Available at <http://archive.archaeology.org/9903/newsbriefs/egypt.html>. [Accessed on 02/04/2015].

Lenhart, A. (2015). *Teens, Social Media & Technology Overview 2015*. [Online]. Available at <http://www.pewinternet.org/files/2015/04/PI_TeensandTech_Update 2015_0409151.pdf />. [Accessed on 23/04/2015].

Lippmann, W. (1920). *Liberty and the News*. New York, N.Y., Harcourt, Brace and Howe.

Lippmann, W. (1965). *Public Opinion*. New York, N.Y., Free Press.

Lippmann, W. (1993). *The Phantom Public*. New Brunswick, N.J., Transaction.

Lohr, S. (2015). *Data-ism. The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else*. London, Oneworld Publications.

Lorenz, M. (2010). *Data-driven journalism: Status and Outlook*. [Online]. Available at <http://www.slideshare.net/mirkolorenz/data-driven-adam>. [Accessed on 03/04/2015].

Lorenz, M. (2010). *Data-driven journalism: What is there to learn?*. [Online]. Available at <http://www.slideshare.net/mirkolorenz/datadriven-journalism-what-is-there-to-learn>. [Accessed on 03/04/2015].

Matsa, K. and Mitchell, A. (2014). *8 Key Takeaways about Social Media and News*. [Online]. Available at <http://www.journalism.org/2014/03/26/8-key-takeaways-about-social-media-and-news/>. [Accessed on 27/05/2015].

Mayfield, A. (2007). *What is Social Media?*. [Online]. Available at <http://www.icross ing.co.uk/fileadmin/uploads/eBooks/What_is_Social_Media_iCrossing_ebook.pdf>. [Accessed on 04/05/2015]

Mazzetti, M. et al. (2015). *SEAL Team 6: A Secret History of Quiet Killings and Blurred Lines*. [Online]. Available at <http://www.nytimes.com/2015/06/07/world/ asia/the-secret-history-of-seal-team-6.html?_r=0>. [Accessed on 06/06/2015].

McCombs, M. E. (1981). The Agenda-Setting Approach. *In*: Nimmo, D. D. and Sanders, K. R. (Ed.). *The Handbook of Political Communication*. Beverly Hills, Calif., Sage.

McCombs, M. E. and Shaw, D. (1972). The Agenda-Setting Function of Mass Media. *In*: *Public Opinion Quarterly*, 36 (2), pp. 176-187.

Meier, K. (2008). *The changing relationship between journalists and their audiences: Drifting together or drifting apart?*. Klaus Meier. [Online]. Available at <http://klaus-

meier.net/blog1/wp-content/uploads/2008/10/meier_iapa_madrid_oct_08.pdf>. [Accessed on 22/05/2015].

Meyer, P. (2002). *Precision Journalism: A Reporter's Introduction to Social Science Methods*. Lanham, Md., Rowman & Littlefield.

Miller, M. and Raisma, M. (2015). *3 ways BBC uses data journalism to increase reader engagement*. [Online]. Available at <http://www.inma.org/blogs/conference/post.cfm/3-ways-bbc-uses-data-journalism-to-increase-reader-engagement>. [Accessed on 15/06/2015].

Mitchell, A. (2013). *News Use across Social Media Platform*. [Online]. Available at <http://www.journalism.org/files/2013/11/News-Use-Across-Social-Media-Platforms1.pdf>. [Accessed on 08/05/2015].

Mitchell, A. (2015). *The Evolving Role of News on Twitter and Facebook*. [Online]. Available at <http://www.journalism.org/files/2015/07/Twitter-and-News-Survey-Report-FINAL2.pdf>. [Accessed on 24/06/2015].

Mitchell, A. and Guskin, E. (2013). *Twitter News Consumers: Young, Mobile and Educated*. [Online]. Available at <http://www.journalism.org/2013/11/04/twitter-news-consumers-young-mobile-and-educated/>. [Accessed on 28/05/2015].

Nixon, R. (2013). *U.S. Postal Service Logging All Mail for Law Enforcement*. [Online]. Available at <http://www.nytimes.com/2013/07/04/us/monitoring-of-snail-mail.html?_r=1>. [Accessed on 12/04/2015].

O'Reilly Radar Team. (2012). *Big data now current perspectives*. Sebastopol, Calif., O'Reilly Media. (Kindle ed.).

Olson, D. R. (2014). *Writing*. [Online]. Available at <http://www.britannica.com/topic/writing>. [Accessed on 26/05/2015].

Open Knowledge Foundation. (n.d.). *The Open Data Handbook*. [Online]. Available at <http://opendatahandbook.org/guide/en/>. [Accessed on 25/05/2015].

Parikh, P. (2014). *How to Lie with Data Visualization*. [Online]. Available at <http://data.heapanalytics.com/how-to-lie-with-data-visualization/>. [Accessed on 27/05/2015].

Patil, D. (2012). *Data Jujitsu the Art of Turning Data into Product*. Sebastopol, Calif., O'Reilly. (Kindle ed.).

Pear Analytics. (2009). *Twitter Study – August 2009*. [Online]. Available at <http://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf>. [Accessed on 30/05/2015].

Pennystocks. (2015). *The Internet in Real-Time*. [Online]. Available at <http://pennystocks.la/internet-in-real-time/>. [Accessed on 15/05/2015].

Pew Research Center. (2007). *A Portrait of "Generation Next*". [Online]. Available at <http://www.people-press.org/2007/01/09/a-portrait-of-generation-next/>. [Accessed on 06/06/2015].

Pew Research Center. (2007). *How Young People View Their Lives, Futures and Politics. A portrait of "Generation Next"*. [Online]. Available at <http://www.people-press.org/files/legacy-pdf/300.pdf>. [Accessed on 23/03/2015].

Pew Research Center. (2012). *Further Decline in Credibility Ratings for Most News Organizations*. [Online]. Available at <http://www.people-press.org/files/2012/08/8-16-2012-Media-Believability1.pdf>. [Accessed on 03/05/2015].

Pew Research Center. (2012). *Trends in News Consumption: 1991-2012. In Changing News Landscape, Even Television is Vulnerable.* [Online]. Available at <http://www.people-press.org/files/legacy-pdf/2012%20News%20Consumption%20Report.pdf>. [Accessed on 26/06/2015].

Pew Research Center. (2015). *Millenials & Political News. Social Media – the Local TV for the Next Generation?*. [Online]. Available at <http://www.journalism.org/files/2015/06/Millennials-and-News-FINAL.pdf>. [Accessed on 08/05/2015].

ProPublica. (2015). *Journalism in the Public Interest*. [Online]. Available at <https://www.propublica.org/>. [Accessed on 29/04/2015].

Remington, A. (2012). *Social science done on deadline: Research chat with ASU's Steve Doig on data journalism*. [Online]. Available at <http://journalistsresource.org/tip-sheets/research/research-chat-steve-doig-data-journalism-social-science-deadline> [Accessed on 27/05/2015].

Resnick, P. et. al. (2014). *RumorLens: A System for Analyzing the Impact of Rumors and Corrections in Social Media*. [Online]. Available at <http://compute-cuj.org/cj-2014/cj2014_session2_paper3.pdf>. [Accessed on 13/06/2015].

Richter, F. (2014). *Infographic: Snapchat More Popular Than Twitter Among Millennials*. Statista Infographics. [Online]. Available at <http://www.statista.com/chart/2570/most-popular-social-apps-among-millennials/>. [Accessed on 01/03/2015].

Robertson, J. (2013). *States' Hospital Data for Sale Puts Privacy in Jeopardy*. [Online]. Bloomberg. Available at <http://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy>. [Accessed 16/05/2015].

Robinson, D. G. and Yu, H. (2012). The New Ambiguity of "Open Government". *59 UCLA Law Review Disc. 178*. [Online]. Available at <http://dx.doi.org/10.2139/ssrn.2012489>. [Accessed on 14/05/2015].

Rodrigues, F. (2015). *Math for journalists. Module 4: Reporting on studies and surveys. Tricky Graphs*. Knight Center for Journalism in the Americas. [Online]. Available at <http://www. journalismcourses.org/>. [Accessed on 12/07/2015].

Rogers, S. (2011). *Data journalism at the Guardian: What is it and how do we do it?*. The Guardian. [Online]. Available at <http://www.theguardian.com/news/datablog/2011/jul/28/data-journalism>. [Accessed on 09/05/2015].

Rogers, S. (2011). *Data journalism broken down: what we do to the data before you see it*. The Guardian. [Online]. Available at <http://www.theguardian.com/news/datablog/2011/apr/07/data-journalism-workflow>. [Accessed on 03/04/2015].

Rogers, S. (2013). *Facts Are Sacred: The Power of Data*. London, Faber and Faber.

Rogers, S. (2015). *Can data journalism be taught?*. [Online]. Simon Rogers. Available at <http://simonrogers.net/2015/01/16/can-data-journalism-be-taught/>. [Accessed on 10/06/2015].

Rogers, S. (2015). *Twitter and data visualisations*. [Online]. Available at <https://prezi.com/ok2ddrtvu8mo/twitter-and-data-visualisations/>. [Accessed on 09/05/ 2015].

Rosen, J. (2008). *PressThink: A Most Useful Definition of Citizen Journalism*. [Online]. Available at <http://archive.pressthink.org/2008/07/14/a_most_useful_d_p.html>. [Accessed on 03/04/2015].

Rosen, J. (2012). *Rosen's Trust Puzzler: What Explains Falling Confidence in the Press?* [Online]. Available at <http://pressthink.org/2012/04/rosens-trust-puzzler-what-explains-falling-confidence-in-the-press/>. [Accessed on 16/06/2015].

SAS Institute Inc. (2015). *Data Visualization: What it is and why it's important*. [Online]. Available at <http://www.sas.com/en_us/insights/big-data/data-visualization.html>. [Accessed on 02/06/2015].

Schrager, A. (2014). *The problem with data journalism*. [Online]. Available at <http://qz.com/189703/the-problem-with-data-journalism/>. [Accessed on 07/06/2015].

Schudson, M. (1995). *The Power of News*. Cambridge, Mass., Harvard University Press.

Schudson, M. (2008). *News and Democratic Society: Past, Present, and Future*. [Online]. Available at < http://www.iasc-culture.org/eNews/2009_10/Schudson_LO.pdf>. [Accessed on 22/06/2015]

Shankland, S. (2009). *The 'Twitter Effect': Possibilities and limits*. [Online]. Available at <http://www.cnet.com/news/the-twitter-effect-possibilities-and-limits/>. [Accessed on 29/05/2015].

Silverman, C. (2014). *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*. Maastricht, European Journalism Centre.

Skowronski, L. (n.d.). *Citizen Journalist "Data Journalism Tip Sheet"*. [Online]. Available at <https://d3n8a8pro7vhmx.cloudfront.net/thecitizenscampaign/pages/161/attachments/original/1364933183/Data_Journalism_Tip_Sheet.pdf?1364933183>. [Accessed on 05/05/2015].

Sky News. (2014). *Sky Youth Poll*. [Online]. Available at <http://survation.com/wp-content/uploads/2014/09/Sky-Youth-Poll-Tables.pdf>. [Accessed on 03/06/2015].

Smiciklas, M. (2012). *The Power of Infographics. Using Pictures to Communicate and Connect with Your Audiences*. Indianapolis, Ind., Que Pub.

Smith, A. (2015). *U.S. Smartphone Use in 2015*. [Online]. Available at <http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/>. [Accessed on 17/05/15].

Smith, N. and Wollan, R. (2011). *The Social Media Management Handbook Everything You Need To Know To Get Social Media Working In Your Business*. Hoboken, N.J., Wiley.

Smola, K. and Sutton, C. (2002). Generational differences: Revisiting generational work values for the new millennium. *In*: *Journal of Organizational Behavior*, 26 (4), pp. 363-382.

Solomon, M. (2015). *2015 Is The Year Of The Millennial Customer: 5 Key Traits These 80 Million Consumers Share*. [Online]. Available at <http://www.forbes.com/sites/micahsolomon/2014/12/29/5-traits-that-define-the-80-million-millennial-customers-coming-your-way/>. [Accessed on 12/05/2015].

Spiller, R. and Weinacht, S. (2014). Datenjournalismus in Deutschland: Eine explorative Untersuchung zu Rollenbildern von Datenjournalisten. *In*: *Publizistik*, 59 (4), pp. 411-433.

Statista. (2015). *Anzahl der aktiven Twitter-Nutzer in ausgewählten Ländern 2013*. [Online]. Available at <http://de.statista.com/statistik/daten/studie/244178/umfrage/aktiven-twitter-nutzer-in-deutschland-und-ausgewaehlten-laendern/>. [Accessed on 30/05/2015].

Statista. (2015). *Twitter: number of monthly active users 2015*. [Online]. Available at <http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>. [Accessed on 01/06/2015].

Statistic Brain. (2015). *Twitter Statistics*. [Online]. Available at <http://www.statistic brain.com/twitter-statistics/>. [Accessed on 25/03/2015].

Stephens, M. (2014). *Beyond News: the future of journalism*. New York, N.Y., Columbia University Press.

Sur, M. (1996). *MIT Research - Brain Processing of Visual Information*. [Online]. Available at <http://newsoffice.mit.edu/1996/visualprocessing>. [Accessed on 22/05/2015].

The Guardian. (2014). *Generation Y takeover: the issues that matter to us and why*. [Online]. Available at <http://www.theguardian.com/lifeandstyle/2014/mar/14/generation-y-takeover-guardian-digital-journalists>. [Accessed on 22/05/2015].

The Pulitzer Prizes. (2013). *Aurora Theater Shootings. #theatershooting*. [Online]. Available at <http://www.pulitzer.org/files/2013/breaking-news-reporting/aurorabreakingnews02.pdf>. [Accessed on 26/03/2015].

Thibodeaux, T. (2015). *5 tips for getting started in data journalism*. [Online]. Available at <http://www.poynter.org/news/media-innovation/147734/5-tips-for-getting-started-in-data-journalism/>. [Accessed on 27/05/2015].

Tufte, E. R. (2001). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn., Graphics Press.

Twitter Help Center. (2014). *The Twitter Glossary*. [Online]. Available at <https://support.twitter.com/groups/50-welcome-to-twitter/topics/204-the-basics/articles/166337-the-twitter-glossary>. [Accessed on 26/04/2015].

Twitter. (2014). *Ellen DeGeneres on Twitter*. [Online]. Available at <https://twitter.com/TheEllenShow/status/440322224407314432/photo/1>. [Accessed on 29/05/2015].

Twitter. (2015). *Company. About.* [Online]. Available at <https://about.twitter.com/company>. [Accessed on 23/03/2015].

United Nations. (2010). *Areas not elsewhere specified (Areas, NES, Bunkers, Special Partners, UN Comtrade)*. [Online]. Available at <http://unstats.un.org/unsd/tradekb/Knowledgebase/Areas-not-elsewhere-specified>. [Accessed on 05/06/2015].

Van Ess, H. (2013). *Driven by Data*. [Online]. Available at <http://www.slideshare.net/searchbistro/driven-by-data-henk-van-ess?qid=d651128d-0174-4f5f-a37d-4c737a2c3b7d&v=default&b=&from_search=1>. [Accessed on 03/05/2015].

von Lucke, J. (2010). *Open Government - Öffnung von Staat und Verwaltung, Gutachten für die Deutsche Telekom AG zur T-City Friedrichshafen*. [Online]. Available at <https://www.zu.de/institute/togi/assets/pdf/TICC-101203-OpenGovernmentData-V1.pdf>. [Accessed on 25/06/2015]

Walter, E. and Gioglio, J. (2015). *The Power of Visual Storytelling: How to Use Visuals, Videos, and Social Media to Market Your Brand*. New York, N.Y., McGraw-Hill Education.

Ward, B. (2014). *Watchdog Culture: Why You Need it, How You Can Build it*. [Online]. Available at <http://www.poynter.org/how-tos/leadership-management/67742/watchdog-culture-why-you-need-it-how-you-can-build-it/>. [Accessed on 16/05/2015].

Ware, C. (2004). *Information Visualization Perception for Design*. San Francisco, Calif., Morgan Kaufman.

Zanchelli, M. and Crucianelli, S. (2012). *Integrating Data Journalism into Newsrooms*. International Center for Journalists. [Online]. Available at <http://www.icfj.org/sites/default/files/integrating%20data%20journalism-english_0.pdf>. [Accessed on 21/06/2015].

ZEIT ONLINE. (2014). *Umfrage: Fast jeder Zweite misstraut den Medien*. [Online]. Available at <http://www.zeit.de/politik/deutschland/2014-12/umfrage-medien-russland-putin-kriegsgefahr>. [Accessed on 03/03/2015].

Images

Figure 1. "*Data-Driven Journalism = Process*". (2010). Available at < http://www.mirkolorenz.com/img/98b52269ae96ac84c263cf258c970a07.png>. [Accessed on 03/04/2015].

Figure 2. "*The inverted pyramid of data journalism*". (2011). Available at < https://onlinejournalismblog.files.wordpress.com/2011/07/inverted-pyramid-of-journalism.jpg>. [Accessed on 16/05/2015].

Figure 3. "*Gathering data: a flow chart of tools and techniques*". (2011). Available at <http://farm7.static.flickr.com/6208/6078887277_5722f1493c.jpg>. [Accessed on 03/04/2015].

Figure 4. "*Workflow - The Guardian*" (2011). Available at <http://www.theguardian.com/news/datablog/2011/apr/07/data-journalism-workflow>. [Accessed on 03/04/2015].

Figure 5. "*Where did you get your news yesterday?*". (2012). Available at <http://www.people-press.org/files/legacy-pdf/2012%20News%20Consumption%20Report.pdf>. [Accessed on 26/06/2015].

Figure 6. "*Should journalists learn to code?*". (2013). Available at <https://pando-assets.s3.amazonaws.com/uploads/2013/10/coderflowchart.jpg>. [Accessed on 18/05/2015].

Figure 7. "*Modes of visualization*". (2015). Available at <http://d396qusza40orc.cloudfront.net/datavisualization/recoded_videos%2F01-13-overview.mpg.aaf4bf202e6111e5b3bcd94ead694dbf.webm>. [Accessed on 12/06/2015]. (Edited by Thomas Schulze).

Figure 8. "*Images versus Text*". In: Smiciklas, M. (2012). *The Power of Infographics. Using Pictures to Communicate and Connect with Your Audiences*. Indianapolis, Ind., Que Pub. (Edited by Thomas Schulze).

Figure 9. "*Diagram of the causes of mortality*". (1858). Available at <http://s3.media.squarespace.com/production/482333/5498857/_V1hky3QMM4k/Sw04NGMgSSI/AAAAAAAAB6s/byqT4FUE5mg/s400/Nightingale-mortality.jpg>. [Accessed on 03/04/2015].

Figure 10. "*French invasion of Russia*". (1869). Available at <http://d.fastcompany.net/multisite_files/fastcompany/imagecache/inline-large/inline/2014/06/3031649-inline-i-minardlg.jpg>. [Accessed on 22/05/2015].

Figure 11. "*Trifecta Checkup*". (2010). Available at <http://junkcharts.typepad.com/.a/6a00d8341e992c53ef0134805aa831970c-pi>. [Accessed on 27/05/2015]. (Edited by Thomas Schulze).

Figure 12. "*8 Hats of Data Visualization Design*". (2012). Available at < http://www.visualisingdata.com/blog/wp-content/uploads/2012/06/HatsProcess.png>. [Accessed on 12/05/2015]. (Edited by Thomas Schulze).

Figure 13. "*Incomplete data*". (2014). Available at <http://cdn1.tnwcdn.com/wp-content/blogs.dir/1/files/2015/05/viz9.jpg>. [Accessed on 27/05/2015].

Figure 14. "*Same data, different y-axis*". (2014). Available at <https://s3.amazonaws.com/heapdatablog/misleading1_yaxis.png>. [Accessed on 27/05/2015].

Figure 15. "*Cumulative graphs*". (2013). Available at <http://qzprod.files.wordpress.com/2013/09/lb_7986_adjusted_bw.png?w=1020&h=620>. [Accessed on 27/05/2015].

Figure 16. "*Ignoring conventions - misleading pie chart*". (2014). Available at <https://s3.amazonaws.com/heapdatablog/misleading3_pie.png>. [Accessed on 27/05/2015].

Figure 17. "*Ignoring conventions*". (2014). Available at <https://s3.amazonaws.com/heapdatablog/misleading3_deaths.jpg>. [Accessed on 27/05/2015].

Figure 18. “*Ignoring conventions - bubble chart*”. (2014). Available at <https://www.devexpress.com/Subscriptions/DXperience/WhatsNew2008v3/i/charts-bubble2d.png>. [Accessed on 27/05/2015].

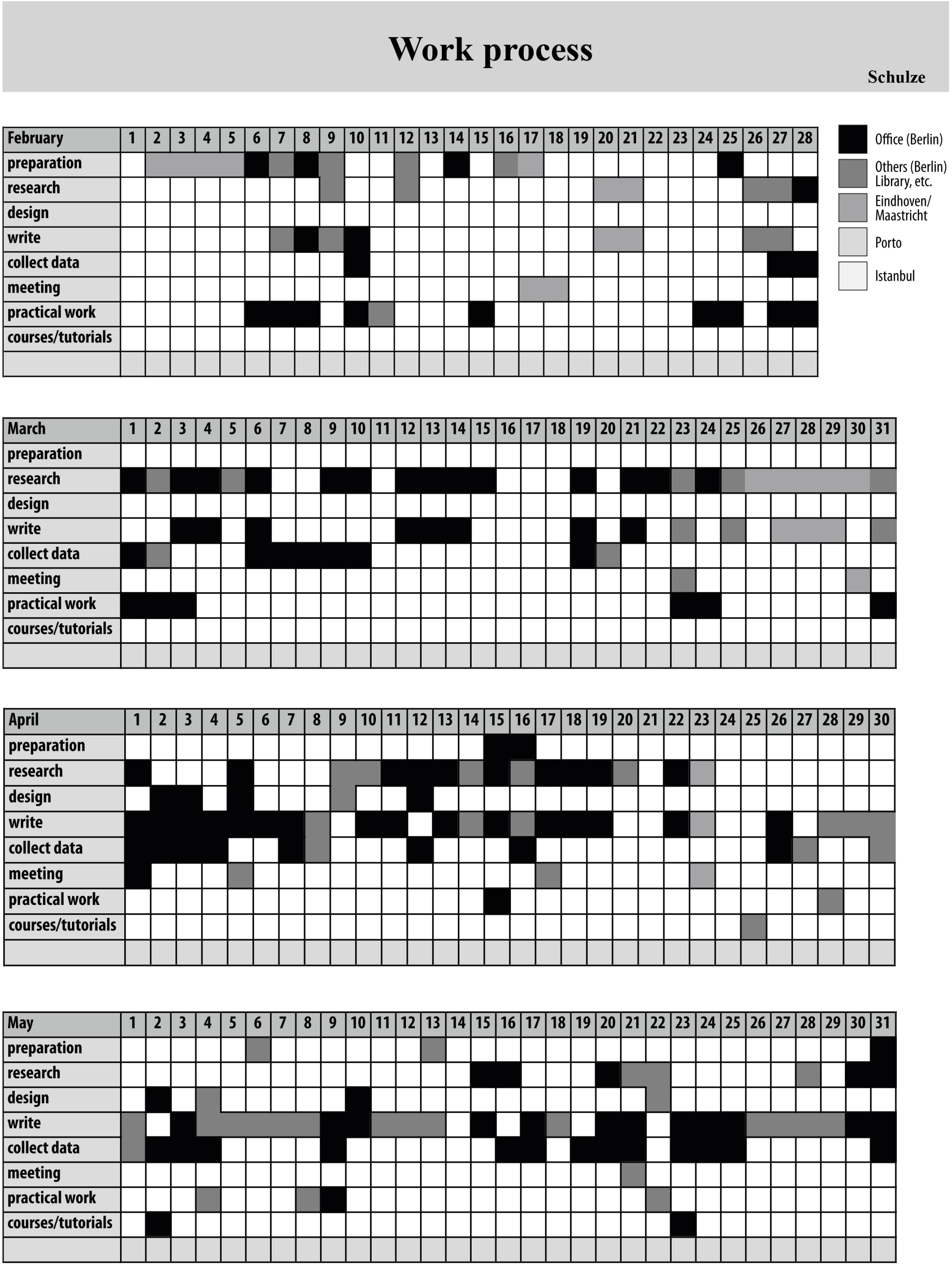
Figure 19. “*Anatomy of a tweet*”. (2014). Available at <https://twitter.com/TheEllenShow/status/440322224407314432/photo/1>. [Accessed on 29/05/2015]. (Edited by Thomas Schulze).

Figure 20. “*Comparison newsrooms*”. (2012). Available at <http://www.icfj.org/sites/default/files/integrating%20data%20journalism-english_0.pdf>. [Accessed on 21/06/2015]. (Edited by Thomas Schulze).

4.2. Appendixes

This part shows processes, images and information that were referred to in the text.

Appendix 1: Work process – Schedule



Work process

Schulze

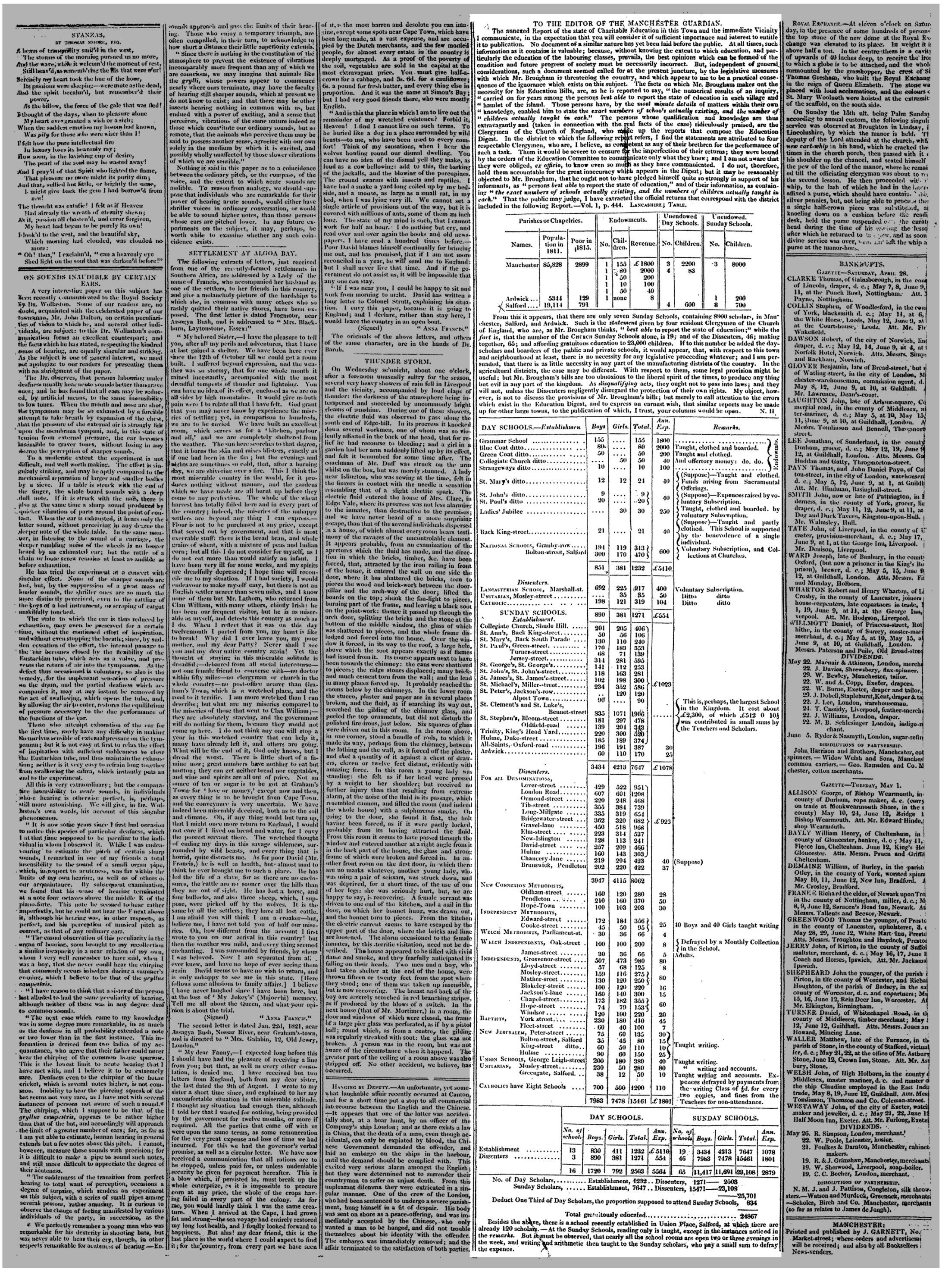
June	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
preparation	▓																												■	■
research																														
design	▓	▓		■														■		■				▒	▒	▒			■	■
write			■			▓	■	▓			■	■	■	▓	▓	■	■	■			■	▓	■							
collect data	▓		■		■		■	■	■	■	■	■	■			■											■	■	■	■
meeting																			▓						▒					
practical work	▓	▓		■														■		■				▒	▒					
courses/tutorials			■	■	■		■			■	■					■	■	■			■							■	■	

July	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
preparation																															
research																										▓					
design				░	░	░	░	░	░	░																			▓	▓	
write	■			░						░				■				■	■					■		▓	■	■	▓	▓	▓
collect data	■	■	■								■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
meeting																															
practical work				░	░	░	░	░	░	░																					
courses/tutorials																							■	■						■	■

August	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
preparation																■	■					■	■	■	■	■	■	■	■	■	■
research	▓					■																									
design			▓					■		▒				■	▓	▓	▓			■		■	■	■	■	▓	▓				
write	▓	■	▓	▓	■	■	■	■	▒	▒	▒		■	■	▓			■	■	■	■						■	■	■		
collect data																															
meeting		▓	▓																						▓	▓					
practical work										▒	■										■	■	■								
courses/tutorials					■	■						■	■																		

Appendix 2: Manchester Guardian – Data Story

Manchester Guardian - Data Story



TO THE EDITOR OF THE MANCHESTER GUARDIAN.

The annexed Report of the state of Charitable Education in this Town and the immediate Vicinity I communicate, in the expectation that you will consider it of sufficient importance and interest to entitle it to publication. No document of a similar nature has yet been laid before the public. At all times, such information as it contains is valuable; because, without knowing the extent to which education, and particularly the education of the labouring classes, prevails, the best opinions which can be formed of the condition and future progress of society must be necessarily incorrect. But, independent of general considerations, such a document seemed called for at the present juncture, by the legislative measures with which Mr. Brougham is threatening the country, and which appear to me to be a practical consequence of the ignorance which exists on this subject. The facts on which Mr. Brougham makes out the necessity for his Education Bills, are, as he is reported to say, "the numerical results of an inquiry, "carried on for years among the persons *best able* to report the state of education in each village and "hamlet of the island. Those persons have, by the *most minute details* of matters within their own "knowledge, enabled him to state the *exact numbers of schools actually existing, and the number of* "*children actually taught in each.*" The persons whose qualification and knowledge are thus extravagantly and (taken in connection with the real facts of the case) ridiculously praised, are the Clergymen of the Church of England, who made up the reports that compose the Education Digest. In the district to which the following report refers, I find the statements are attributed to four respectable Clergymen, who are, I believe, as competent as any of their brethren for the performance of such a task. Them it would be severe to censure for the imperfection of their returns; they were bound by the orders of the Education Committee to communicate only what they knew; and I am not aware that they were obliged, *ex officio*, to know even so much as they have communicated. I do not, therefore, hold them accountable for the great inaccuracy which appears in the Digest; but it may be reasonably objected to Mr. Brougham, that he ought not to have pledged himself quite so strongly in support of his informants, as "persons *best able* to report the state of education," and of their information, as containing "*the exact numbers of schools actually existing, and the numbers of children actually taught in each.*" That the public may judge, I have extracted the official returns that correspond with the district included in the following Report.—Vol. 1, p. 444. LANCASHIRE; TABLE.

Parishes or Chapelries.			Endowments.			Unendowed Day Schools.		Unendowed. Sunday Schools.	
Names.	Population in 1811.	Poor in 1815.	No.	Children.	Revenue.	No.	Children.	No.	Children.
Manchester	85,828	2899	1	155	£1800	3	2200	3	8000
			1	80	2000	4	83		
			1	50	200				
			1	10	100				
			1	50	40				
Ardwick ...	5344	129	1	none	8			1	200
Salford	19,114	791				4	600	8	700

From this it appears, that there are only seven Sunday Schools, containing 8900 scholars, in Manchester, Salford, and Ardwick. Such is the *statement* given by four resident Clergymen of the Church of England, who are, as Mr. Brougham thinks, "*best able* to report the state of education;" while the *fact* is, that the number of the CHURCH Sunday Schools alone, is 19; and of the Dissenters, 46; making together, 65; and affording gratuitous education to 23,000 children. If to this number be added the day-scholars and boarders of the public and private schools, it would appear, that, with respect to this town and neighbourhood at least, there is no necessity for any legislative proceeding whatever; and I am persuaded, that there is as little necessity in any part of the manufacturing districts of the country. In the agricultural districts, the case may be different. With respect to them, some legal provision might be useful; but Mr. Brougham's bills are too obnoxious to the liberal spirit of the times, to produce any thing but evil in any part of the kingdom. As *disqualifying* acts, they ought not to pass into laws; and they will not, unless the Dissenters negligently disregard the protection of their own rights. My object, however, is not to discuss the provisions of Mr. Brougham's bills; but merely to call attention to the errors which exist in the Education Digest, and to express an earnest wish, that similar reports may be made up for other large towns, to the publication of which, I trust, your columns would be open. N. H.

DAY SCHOOLS.—*Establishmen*	*Boys*	*Girls.*	*Total.*	*Ann. Exp.*	*Remarks.*
Grammar School	155	....	155	1800	.... Endowmts.
Blue Coat ditto..................	80	....	80	2000	Taught, clothed and boarded.
Green Coat ditto..................	50	....	50	200	Taught and clothed.
Collegiate Church ditto..........	....	50	50	40	And offertory money: do. do.
Strangeways ditto..................	10	....	10	100	....
St. Mary's ditto..................	12	12	24	40	(Suppose)—Taught and clothed. Funds arising from Sacramental Offerings.
St. John's ditto..................	9	....	9	40	(Suppose)—Expences raised by voluntary Subscription.
St. Paul's ditto	20	....	20		
Ladies' Jubilee	....	30	30	250	Taught, clothed and boarded. by voluntary Subscription.
Back King-street..................	21	....	21	40	(Suppose)—Taught and partly clothed. This School is supported by the benevolence of a single individual.
NATIONAL SCHOOLS, Granby-row....	194	119	313	600	Voluntary Subscription, and Collections at Churches.
Bolton-street, Salford	300	170	470		
	851	381	1232	£5110	
Dissenters.					
LANCASTERIAN SCHOOL, Marshall-st.	692	225	917	400	Voluntary Subscription.
UNITARIAN, Mosley-street	....	35	35	50	Ditto ditto
CATHOLIC............................	198	121	319	104	Ditto ditto
	890	381	1271	£554	

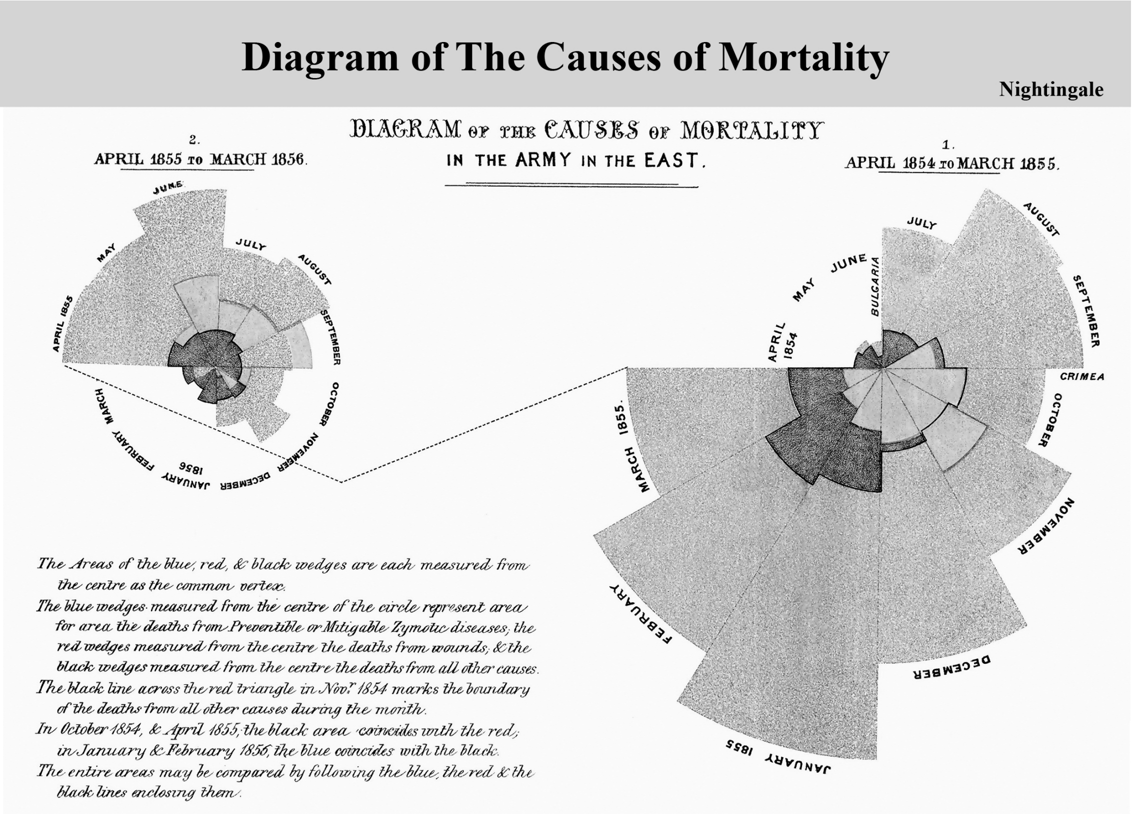
SUNDAY SCHOOLS. *Establishment.*					
Collegiate Church, Shude Hill.	201	205	406	£1023	
St. Ann's, Back King-street........	50	56	106		
St. Mary's, Back South Parade	130	110	240		
St. Paul's, Green-street.	170	183	353		
Turner-street	68	71	139		
Jersey-street.	314	281	595		
St. George's, St. George's..........	141	112	253		
St. John's, St. John's-street........	118	163	281		
St. James's, St. James's-street......	102	198	300		
St. Michael's, Miller-street........	234	352	586		
St. Peter's, Jackson's-row..........	..	120	120		
Alport Town..........	90	....	90		
St. Clement's and St. Luke's, Bennet-street	835	1071	1906	......	This is, perhaps, the largest School in the Kingdom. It cost about £2,300, of which £512 0 10½ was contributed in small sums by the Teachers and Scholars.
St. Stephen's, Bloom-street	181	297	478		
Oldfield-road	139	204	343		
Trinity, King's Head Yard..........	220	300	520		
Hulme, Duke-street	185	189	374		
All-Saints, Oxford-road	196	191	387	30	
Ardwick..............................	60	110	170	25	
	3434	4213	7647	£1078	
Dissenters.					
FOR ALL DENOMINATIONS,					
Lever-street	429	522	951	£923	
London Road........	607	601	1208		
Ormond-street	220	248	468		
Tib-street	355	384	739		
Long-Millgate	335	319	654		
Bridgewater-street	362	320	682		
Gravel-lane	450	518	968		
Elm-street	223	314	537		
New-Islington	128	113	241		
David-street	257	209	466		
Hulme	160	143	303		
Chancery-lane	219	204	423	40	(Suppose)
Brunswick, Pendleton	202	220	422	37	
	3947	4115	8062		
NEW CONNEXION METHODISTS,					
Oldham-street	160	120	280	28	
Pendleton	210	160	370	50	
Hope-Town	100	103	203	30	
INDEPENDENT METHODISTS,					
Edward-street	172	184	356	25	40 Boys and 40 Girls taught writing
Cooke-street	45	50	95		
WELCH METHODISTS, Parliament-st.	30	36	66	4	
WELCH INDEPENDENTS, Oak-street .	100	100	200	8	Defrayed by a Monthly Collection in the School.
James-street	30	36	66	5	Adults.
INDEPENDENTS, Grosvenor-street	507	473	980	80	
Lloyd-street	57	68	125	8	
Mosley-street........	159	116	275	80	
Mather-sreet	130	120	250		
Blakeley-street	100	120	220	16	
Jackson's-lane	160	140	300	15	
Chapel-street........	173	182	355	60	
Hope-street	74	79	153		
Windsor	120	100	220	20	
BAPTISTS, York street....	230	180	410	45	
Fleet-street	60	40	100	7	
NEW JERUSALEM, Peter-street	75	60	135	30	Taught writing.
Bolton-street, Salford	35	45	80	15	
King-street ditto..	60	50	110	10	
Hulme	90	60	150	25	
UNION SCHOOLS, George Leigh-street	200	180	380	40	Taught writing.
UNITARIAN, Mosley-street........	230	50	280	80	" writing and accounts.
Greengate, Salford. ..	38	12	50	10	Taught writing and accounts. Expences defrayed by payments from the writing Class of ½d. for every two copies, and fines from the Teachers for non-attendance.
CATHOLICS have Eight Schools	700	500	1200	110	
	7983	7478	15461	£1801	

	DAY SCHOOLS.					SUNDAY SCHOOLS.				
	No. of schools	*Boys.*	*Girls.*	*Total.*	*Ann. Exp.*	*No. of schools*	*Boys.*	*Girls.*	*Total.*	*Ann. Exp.*
Establishment	13	830	411	1232	£5110	19	3434	4213	7647	1078
Dissenters	3	890	381	1271	554	46	7983	7478	15461	1801
	16	1720	792	2503	5564	65	11,417	11,691	23,108	2879

No. of Day Scholars,.........	Establishment, 1232..	Dissenters, 1271——	2503
Sunday Scholars, ...	Establishment, 7647..	Dissenters, 15471——	23,108
			——25,701
Deduct One Third of Day Scholars, the proportion supposed to attend Sunday Schools,			834
Total gratuitously educated............................			24867

Besides the above, there is a school recently established in Union Place, Salford, at which there are already 120 scholars.—At the Sunday Schools, reading only is taught, except in the instances noticed in the *remarks*. But it must be observed, that nearly all the school rooms are open two or three evenings in the week, and writing and arithmetic then taught to the Sunday scholars, who pay a small sum to defray the expence.

Appendix 3: Nightingale – Diagram of the causes of mortality



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

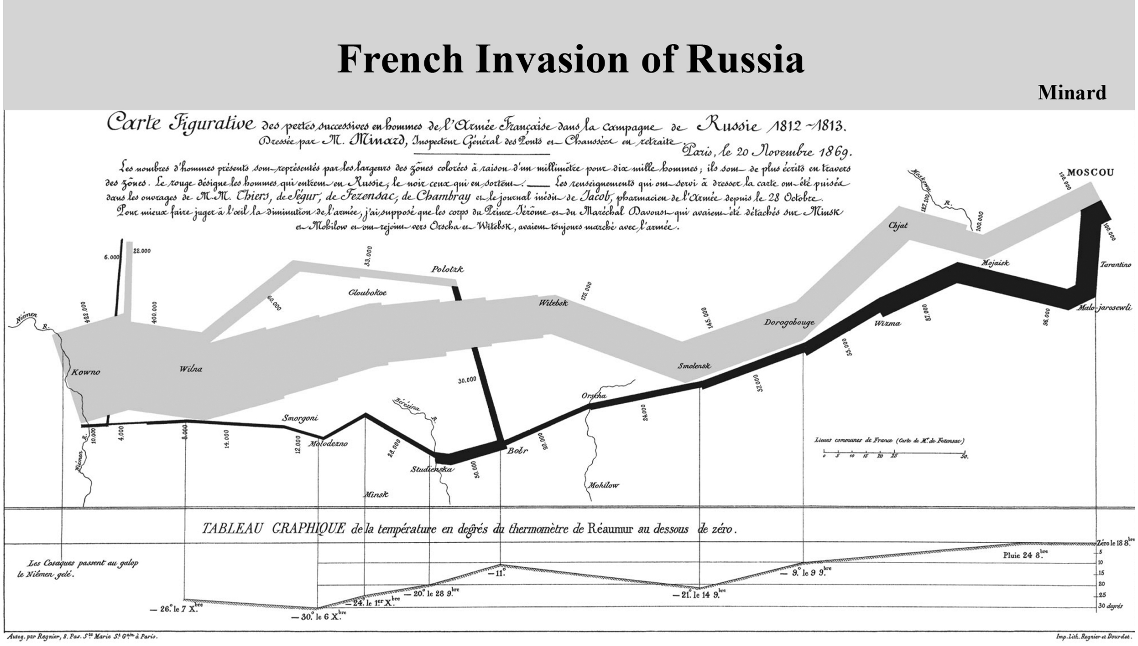
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Appendix 4: Minard – French invasion of Russia



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

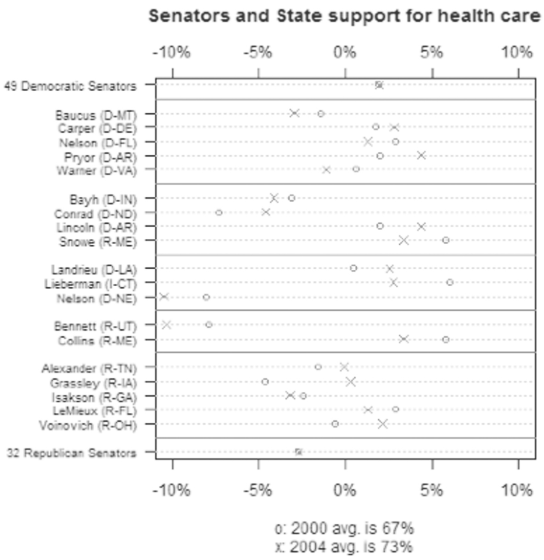
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

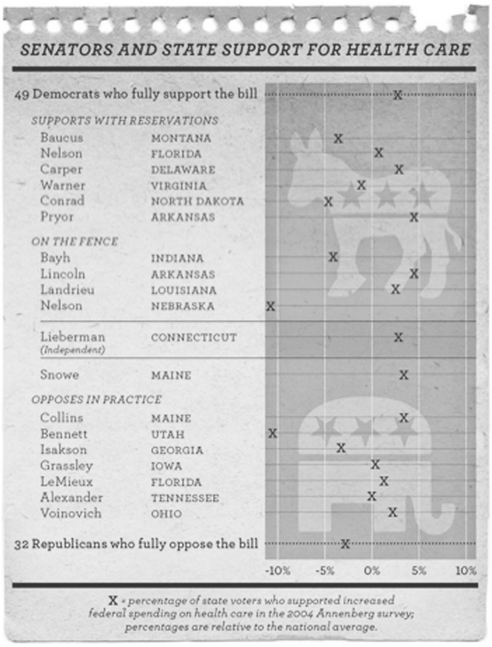
Appendix 5: Practical example - Trifecta Check up

As mentioned when describing the *Trifecta Checkup*, Fung compiled a list of nine changes that were made, in order to make a publish-worthy chart for the New York Times. The image is presented in grayscales and can be found in original color on http://tinyurl.com/pntdw6h.

Trifecta Checkup - Example

Gelman & NYT





1. The data from 2000 was removed and only the 2004 data was relied on.
2. The amount of groups of senators was reduced from seven to five and Senator Lieberman was clearly separated to the two other parties.
3. The senators were re-ordered and classified into Democrats, Independent and Republicans.
4. Annotations that explain the groups of senators were added.
5. The scale labels on top of the of the chart were removed and only the bottom label was kept.
6. The "text area" was increased and takes now half of the chart.
7. White gridlines were added.
8. The background of the chart was colored and the cross marks were put in black. Additionally, yellow color was introduced which was originally also colored blue.

9. As a result, of the clear explanation in the created legend, the references to the national average were removed.

Fung sees several good features in the original chart that was left unmodified, such as the scale of 10% although the highest percentage is 6%. This gives the audience a better impression of the relation of the numbers and helps to judge them. Another helpful feature that was kept are the subdued horizontal gridlines, which help the reader to orient within the chart. Finally, retaining the title of the chart, to summarize the findings and the horizontal scale labels at a 5% interval supports the readability of the chart. Moreover, Fung sees in the points 2, 3 and 4 a definite improvement of the visualization. He also approves 1 (removal of 2000 data) since the data of both years is highly correlated and 5 (removal of top scale) because of the added gridlines in 7.

Debatable points from Fung's point of view are the points 8 (use of different marks and background) and 9 (removal of the "73%" in the legend). He argues that by leaving out the national average of 73%, the audience might get confused and need to understand relative versus absolute values. As the most debatable point, he sees 6 (decrease of the chart itself) because the data-ink ratio has been massively reduced.

Appendix 6: Types of commonly used visualization

Bar charts

These are frequently used to compare a set of numeric values. To create a bar chart you need at lest two columns or rows containing data series and corresponding labels. It can be used for positive and negative values.

Line charts

Line graphs are useful for visualizing continuous change over time. Most often they are used for time-series data, with the time labels in the x-axis and the values in the y-axis. Line charts work well for changes over time spans such as months, quarters, or fiscal years. For example: the house price index.

Pie chart

Pie charts are useful when the aim is to show how the size of values within a series relates to the sum, i.e. with individual categories the total is made up of. For this type of chart data has to be arranged in one column or row. Values cannot be negative.

Scatter plot

Scatter plots let you show relationships between numeric values in several data series. They are often used comparing values such as scientific or statistical data. You can use it for example to compare pay versus performance.

Bubble chart

A bubble chart is a popular way to display a set of numeric values. They are useful if the difference between the data values graphed is very large. The area of the bubbles represents the value. Bubble charts can be represented in a scatter plot with the first two values representing the position of the bubbles on the x and taxes of the chart, and the third value representing the size of the bubble. However, bubble charts can also just show you one value, by means of the size of the bubble and the circles can be packed together tightly to save space.

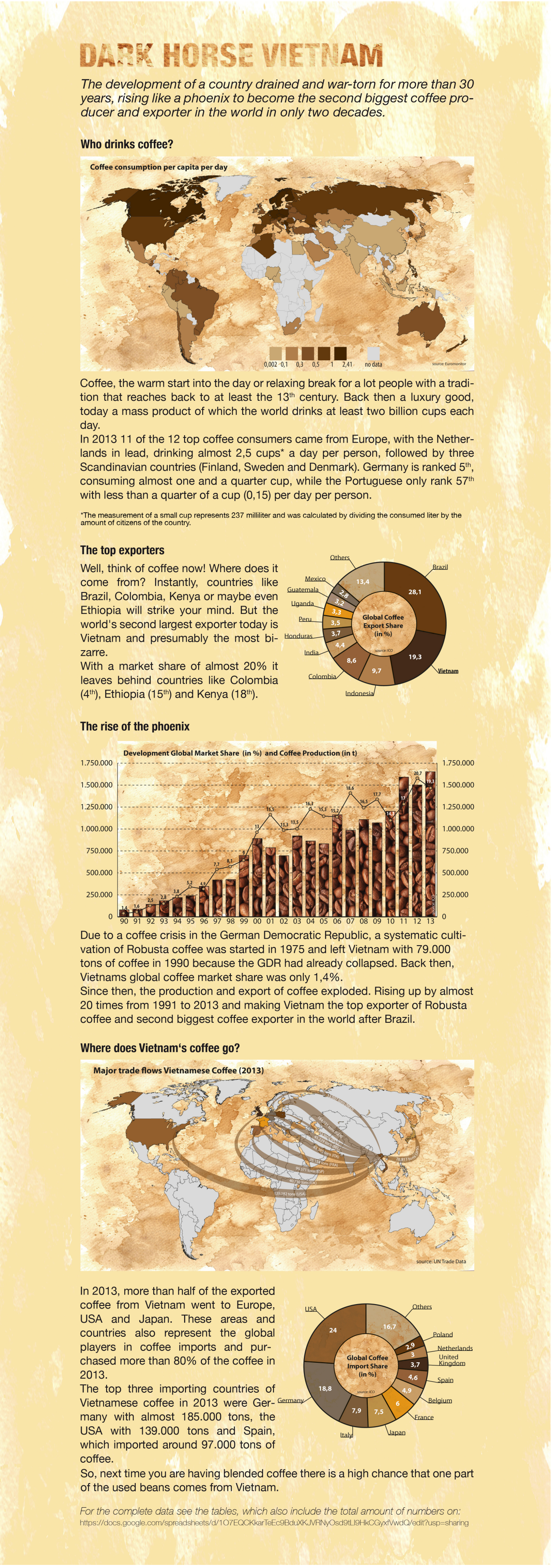
Words cloud

Tools like “*Many Eyes*” and “*Wordle*” allow you to import text to reveal word frequencies to create a word cloud, also known as tag cloud, you need to import either one or two text files and the tool takes care of the rest. Tag clouds can be a fun way to look at text, but there can be problems such as meaningless words like “the” can be over-emphasized.

Appendix 7: Dark Horse Vietnam – Article

This attachment shows the printed version of the example of use. For a high-resolution version of the article, see:

https://drive.google.com/file/d/0B8jjkb5sAYYqSlBxOTNKa0lzc3M/view?usp=sharing



DARK HORSE VIETNAM

The development of a country drained and war-torn for more than 30 years, rising like a phoenix to become the second biggest coffee producer and exporter in the world in only two decades.

Who drinks coffee?

Coffee consumption per capita per day
0,002
0,1
0,3
0,5
1
2,41
no data
source: Euromonitor

Coffee, the warm start into the day or relaxing break for a lot people with a tradition that reaches back to at least the 13th century. Back then a luxury good, today a mass product of which the world drinks at least two billion cups each day.

In 2013 11 of the 12 top coffee consumers came from Europe, with the Netherlands in lead, drinking almost 2,5 cups* a day per person, followed by three Scandinavian countries (Finland, Sweden and Denmark). Germany is ranked 5th, consuming almost one and a quarter cup, while the Portuguese only rank 57th with less than a quarter of a cup (0,15) per day per person.

*The measurement of a small cup represents 237 milliliter and was calculated by dividing the consumed liter by the amount of citizens of the country.

The top exporters

Well, think of coffee now! Where does it come from? Instantly, countries like Brazil, Colombia, Kenya or maybe even Ethiopia will strike your mind. But the world's second largest exporter today is Vietnam and presumably the most bizarre.

With a market share of almost 20% it leaves behind countries like Colombia (4th), Ethiopia (15th) and Kenya (18th).

The rise of the phoenix

Due to a coffee crisis in the German Democratic Republic, a systematic cultivation of Robusta coffee was started in 1975 and left Vietnam with 79.000 tons of coffee in 1990 because the GDR had already collapsed. Back then, Vietnams global coffee market share was only 1,4%.

Since then, the production and export of coffee exploded. Rising up by almost 20 times from 1991 to 2013 and making Vietnam the top exporter of Robusta coffee and second biggest coffee exporter in the world after Brazil.

Where does Vietnam's coffee go?

In 2013, more than half of the exported coffee from Vietnam went to Europe, USA and Japan. These areas and countries also represent the global players in coffee imports and purchased more than 80% of the coffee in 2013.

The top three importing countries of Vietnamese coffee in 2013 were Germany with almost 185.000 tons, the USA with 139.000 tons and Spain, which imported around 97.000 tons of coffee.

So, next time you are having blended coffee there is a high chance that one part of the used beans comes from Vietnam.

For the complete data see the tables, which also include the total amount of numbers on:
https://docs.google.com/spreadsheets/d/1O7EQCKkarTeEc9BduXKJVRNyOsd9tLI9HkCGyxfVwdQ/edit?usp=sharing

Appendix 8: Twitter Glossary

The glossary refers to the Help Center of the Twitter (Twitter Help Center, 2014) and explains all terms used in connection to the use of Twitter.

@: The @ sign is used to call out usernames in Tweets: "Hello @twitter!" People will use your @username to mention you in Tweets, send you a message or link to your profile.

@username: A username is how you're identified on Twitter, and is always preceded immediately by the @ symbol. For instance, Katy Perry is @katyperry.

Alerts: Twitter Alerts enable public safety agencies to inform people during emergencies by highlighting critical time-sensitive content with notifications and a unique look.

Bio: Your bio is a short (up to 160 characters) personal description that appears in your profile that serves to characterize your persona on Twitter.

Block: If you block a Twitter user, that account will be unable to follow you or add you to their Twitter lists, and you will not receive a notification if they mention you in a Tweet.

Bug: An internal error in our site code and functionality. We find and fix them all the time (nobody's perfect). If you see one, point it out to @Support by sending us a message.

Cashtag: A cashtag is a company ticker symbol preceded by the U.S. dollar sign, e.g. $TWTR. When you click on a cashtag, you'll see other Tweets mentioning that same ticker symbol.

Deactivation: If you deactivate your account, it goes into a queue for permanent deletion from Twitter in 30 days. You may reactivate your account within the 30 day grace period.

Direct Messages: Direct Messages are private messages sent from one Twitter user to another Twitter users. You can use Direct Messages for one-on-one private conversations, or between groups of users.

Discover: This feature surfaces personalized content tailored to your interests.

Favorite: Favoriting a Tweet indicates that you liked a specific Tweet. You can find all of your favorite Tweets by clicking on the favorites link on your profile page.

Tap the star icon to favorite a Tweet and the author will see that you liked it.

Follow: Subscribing to a Twitter account is called “following.” To start following, click the Follow button next to the user name or on their profile page to see their Tweets as soon as they post something new. Anyone on Twitter can follow or unfollow anyone else at any time, with the exception of blocked accounts. See "block."

A follow is the result of someone following your Twitter account. You can see how many follows (or followers) you have from your Twitter profile.

Follow button: Click the Follow button to follow (or unfollow) anyone on Twitter at any time. When you follow someone, you will see their Tweets in your Home stream.

Follow count: This count reflects how many people you follow and how many follow you; these numbers are found on your Twitter profile.

Follower: A follower is another Twitter user who has followed you to receive your Tweets in their Home stream.

Geolocation, geotagging: Adding a location to your tweet (a geolocation or geotag) tells those who see your Tweet where you were when you posted that Tweet.

Hacking: Gaining unauthorized access to an account via phishing, password guessing, or session stealing. Usually this is followed by unauthorized posts from the account. Hacked accounts are sometimes referred to as "compromised." Click here if you've been hacked. Read more about how to keep your account safe.

Hashflag: A hashflag is a specific series of letters immediately preceded by the # sign which generates an icon on Twitter such as a national flag or another small image.

Hashtag, #: A hashtag is any word or phrase immediately preceded by the # symbol. When you click on a hashtag, you'll see other Tweets containing the same keyword or topic.

Header photo: Your personal image that you upload, which appears at the top of your profile.

Home: Home is your real-time stream of Tweets from those you follow.

Impersonation: Online impersonation (pretending to be someone you're not) that is intended to deceive is prohibited under the Twitter Rules. Parody accounts are allowed. See "parody."

List: From your own account, you can create a group list of other Twitter users by topic or interest (e.g., a list of friends, coworkers, celebrities, athletes). Twitter lists also contain a timeline of Tweets from the specific users that were added to the list, offering you a way to follow individual accounts as a group on Twitter.

Mention: Mentioning other users in your Tweet by including the @ sign followed directly by their username is called a “mention.” Also refers to Tweets in which your @username was included.

Notifications, notifications: The Notifications timeline displays your interactions with other Twitter users, like mentions, favorites, Retweets and who has recently followed you. If you request it, we send notifications to you via SMS or through the Twitter for iPhone or Twitter for Android apps.

Parody: You can create parody accounts on Twitter to spoof or make fun of something in jest, as well as commentary and fan accounts. These accounts must disclose that they are parody, fan or commentary accounts in order to comply with our strict policy against impersonation. See "impersonation."

Pinned Tweets: You can pin a Tweet to the top of your profile page to keep something important to you above the flow of time-ordered Tweets.

Profile: Your profile displays information you choose to share publicly, as well as all of the Tweets you've posted. Your profile along with your @username identify you on Twitter.

Profile photo: Your personal image found under the Me icon. It's also the picture that appears next to each of your Tweets.

Promoted Accounts: Promoted Accounts present suggested accounts you might want to follow as promoted by our advertisers. These appear in your Home timeline, and via Who to Follow, search results and elsewhere on the platform.

Promoted Trends: Promoted Trends display time-, context-, and event-sensitive trends promoted by our advertisers. These appear at the top of the Trending Topics list on Twitter and elsewhere on the platform, and are clearly marked as "Promoted."

Promoted Tweets: Promoted Tweets are Tweets that are paid for by our advertisers. These appear in your Home timeline, at the top of search results on Twitter and elsewhere on the platform, and are clearly marked as "Promoted."

Protected/private accounts: Twitter accounts are public by default. Choosing to protect your account means that your Tweets will only be seen by approved followers and will not appear in search.

Reply: A response to another user's Tweet that begins with the @username of the person you're replying to is known as a reply. Reply by clicking the "reply" button next to the Tweet you'd like to respond to.

Retweet: A Tweet that you forward to your followers is known as a Retweet. Often used to pass along news or other valuable discoveries on Twitter, Retweets always retain original attribution.

The act of sharing another user's Tweet to all of your followers by clicking on the Retweet button.

Timeline: A timeline is a real-time stream of Tweets. Your Home stream, for instance, is where you see all the Tweets shared by your friends and other people you follow.

Timestamp: The date and time a Tweet was posted to Twitter. A Tweet's timestamp can be found in grey text in the detail view of any Tweet.

Top Tweets: Tweets determined by a Twitter algorithm to be the most popular or resonant on Twitter at any given time. Read more about Top Tweets.

Trends: A Trend is a topic or hashtag determined algorithmically to be one of the most popular on Twitter at that moment. You can choose to tailor Trends based on your location and who you follow.

Tweet, Tweetable: A Tweet may contain photos, videos, links and up to 140 characters of text.

Tweet: The act of sending a Tweet. Tweets get shown in Twitter timelines or are embedded in websites and blogs.

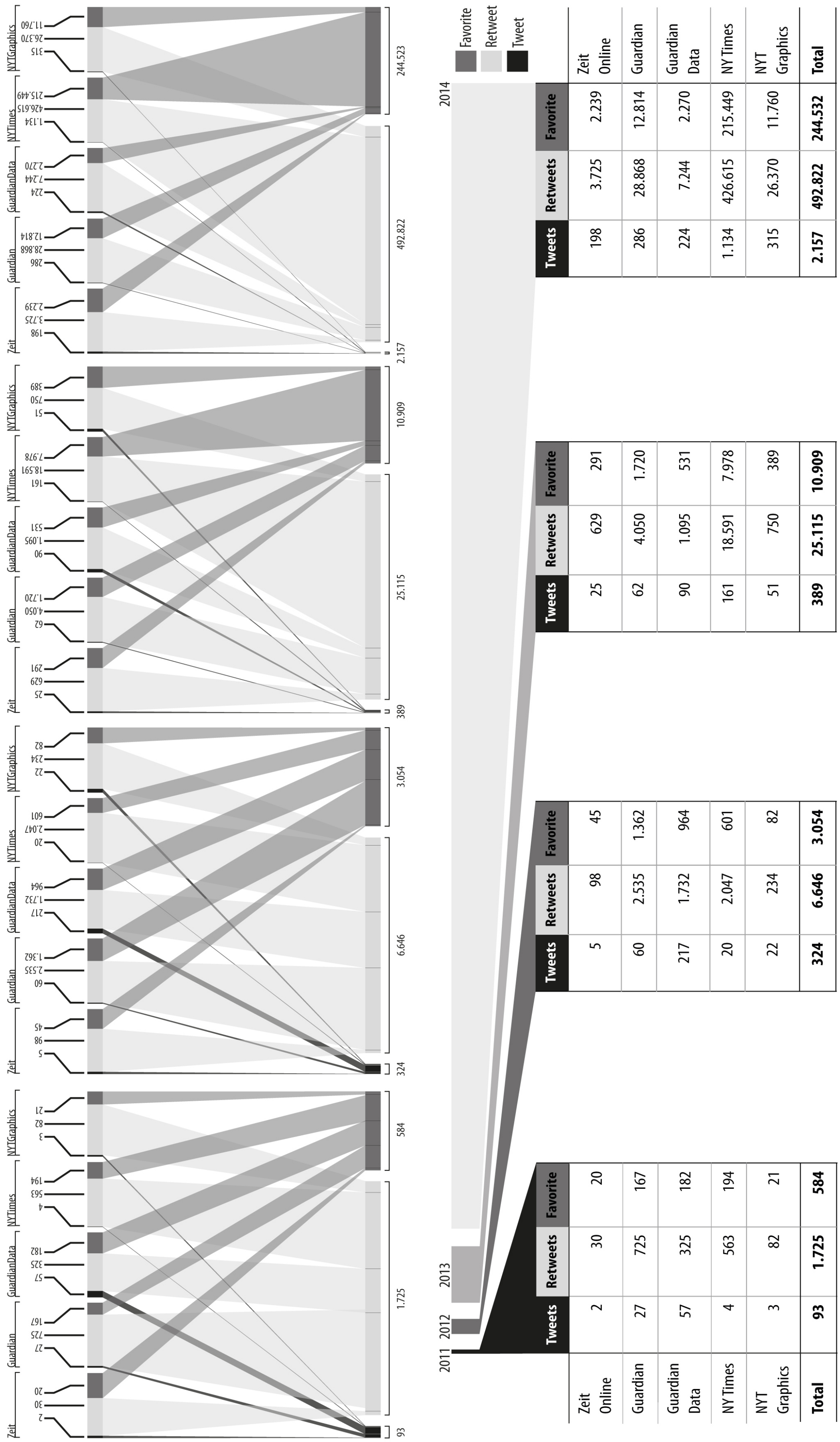
Tweet button: Anyone can add a Tweet button to their website. Clicking this button lets Twitter users post a Tweet with a link to that site. Learn how to add the Tweet button to your website here.

URL, URLs: A URL (Uniform Resource Locator) is a web address that points to a unique page on the Internet.

Verification: A process whereby a Twitter account receives a blue check icon to indicate that the creator of these Tweets is a legitimate source. Verified users include public figures and those who may have experienced identity confusion on Twitter.

Who to follow: Who to follow is an automated list of recommended accounts we think you might find interesting, based on the types of accounts you already follow and who those people follow.

Appendix 9: Visualization – Case Study



2011

	Tweets	Retweets	Favorite
Zeit Online	2	30	20
Guardian	27	725	167
Guardian Data	57	325	182
NY Times	4	563	194
NYT Graphics	3	82	21
Total	**93**	**1.725**	**584**

2012

	Tweets	Retweets	Favorite
Zeit Online	5	98	45
Guardian	60	2.535	1.362
Guardian Data	217	1.732	964
NY Times	20	2.047	601
NYT Graphics	22	234	82
Total	**324**	**6.646**	**3.054**

2013

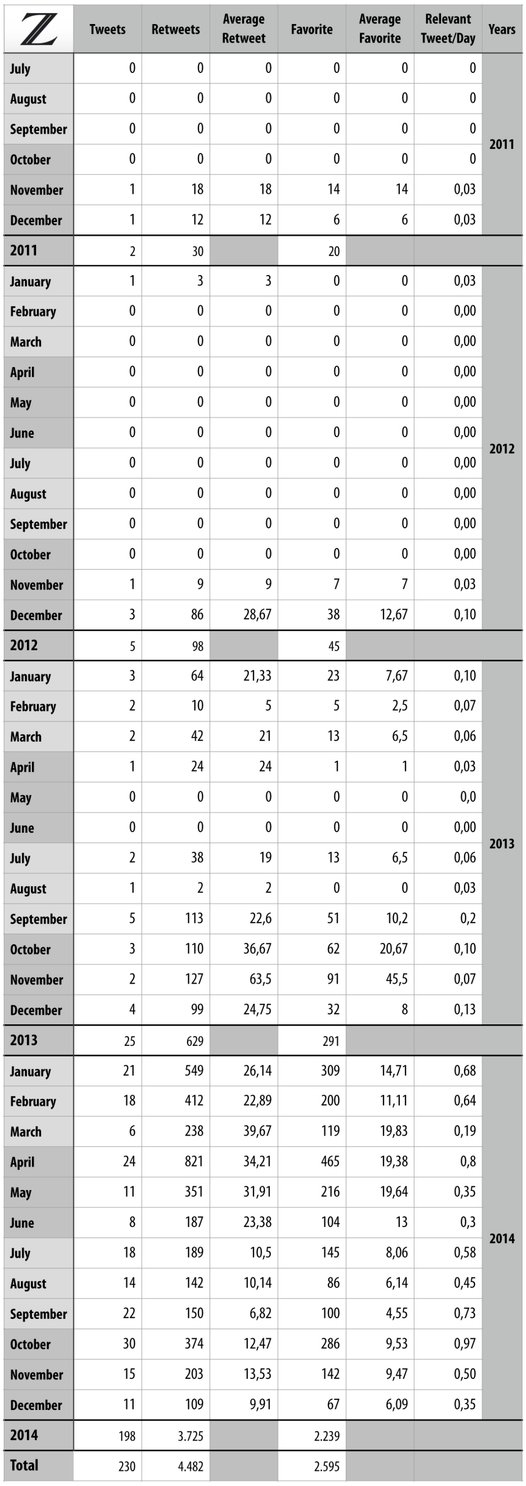
	Tweets	Retweets	Favorite
Zeit Online	25	629	291
Guardian	62	4.050	1.720
Guardian Data	90	1.095	531
NY Times	161	18.591	7.978
NYT Graphics	51	750	389
Total	**389**	**25.115**	**10.909**

2014

	Tweets	Retweets	Favorite
Zeit Online	198	3.725	2.239
Guardian	286	28.868	12.814
Guardian Data	224	7.244	2.270
NY Times	1.134	426.615	215.449
NYT Graphics	315	26.370	11.760
Total	**2.157**	**492.822**	**244.532**

Appendix 10: Dataset - Die ZEIT

Die ZEIT



Z	Tweets	Retweets	Average Retweet	Favorite	Average Favorite	Relevant Tweet/Day	Years
July	0	0	0	0	0	0	2011
August	0	0	0	0	0	0	
September	0	0	0	0	0	0	
October	0	0	0	0	0	0	
November	1	18	18	14	14	0,03	
December	1	12	12	6	6	0,03	
2011	2	30		20			
January	1	3	3	0	0	0,03	2012
February	0	0	0	0	0	0,00	
March	0	0	0	0	0	0,00	
April	0	0	0	0	0	0,00	
May	0	0	0	0	0	0,00	
June	0	0	0	0	0	0,00	
July	0	0	0	0	0	0,00	
August	0	0	0	0	0	0,00	
September	0	0	0	0	0	0,00	
October	0	0	0	0	0	0,00	
November	1	9	9	7	7	0,03	
December	3	86	28,67	38	12,67	0,10	
2012	5	98		45			
January	3	64	21,33	23	7,67	0,10	2013
February	2	10	5	5	2,5	0,07	
March	2	42	21	13	6,5	0,06	
April	1	24	24	1	1	0,03	
May	0	0	0	0	0	0,0	
June	0	0	0	0	0	0,00	
July	2	38	19	13	6,5	0,06	
August	1	2	2	0	0	0,03	
September	5	113	22,6	51	10,2	0,2	
October	3	110	36,67	62	20,67	0,10	
November	2	127	63,5	91	45,5	0,07	
December	4	99	24,75	32	8	0,13	
2013	25	629		291			
January	21	549	26,14	309	14,71	0,68	2014
February	18	412	22,89	200	11,11	0,64	
March	6	238	39,67	119	19,83	0,19	
April	24	821	34,21	465	19,38	0,8	
May	11	351	31,91	216	19,64	0,35	
June	8	187	23,38	104	13	0,3	
July	18	189	10,5	145	8,06	0,58	
August	14	142	10,14	86	6,14	0,45	
September	22	150	6,82	100	4,55	0,73	
October	30	374	12,47	286	9,53	0,97	
November	15	203	13,53	142	9,47	0,50	
December	11	109	9,91	67	6,09	0,35	
2014	198	3.725		2.239			
Total	230	4.482		2.595			

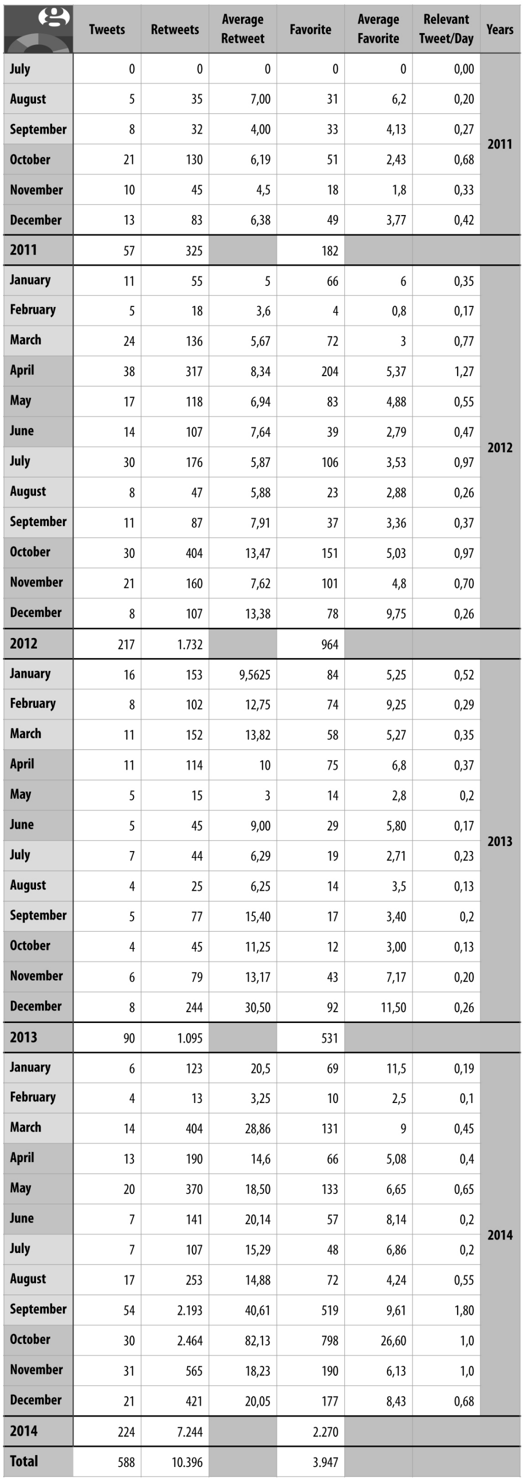
Appendix 11: Dataset - The Guardian

The Guardian

	Tweets	Retweets	Average Retweet	Favorite	Average Favorite	Relevant Tweet/Day	Years
July	0	0	0	0	0	0,00	2011
August	3	98	32,67	12	4	0,10	
September	6	175	29,17	36	6	0,20	
October	4	104	26	19	4,75	0,13	
November	10	285	28,5	74	7,4	0,33	
December	4	63	15,75	26	6,5	0,13	
2011	27	725		167			
January	6	252	42	144	24	0,19	2012
February	0	0	0	0	0	0	
March	7	348	49,71	140	20	0,23	
April	7	209	29,86	88	12,57	0,23	
May	6	174	29	212	35,33	0,19	
June	5	208	41,6	86	17,2	0,17	
July	3	116	38,67	36	12	0,10	
August	6	277	46,17	224	37,33	0,19	
September	2	24	12	9	4,5	0,07	
October	9	446	49,56	221	24,56	0,29	
November	7	414	59,14	163	23,3	0,23	
December	2	67	33,5	39	19,5	0,06	
2012	60	2.535		1362			
January	8	252	31,5	91	11,38	0,26	2013
February	1	166	166	38	38	0,04	
March	4	324	81	155	38,75	0,13	
April	2	22	11	7	3,5	0,07	
May	2	87	43,5	41	20,5	0,1	
June	3	115	38,33	35	11,67	0,10	
July	5	214	42,8	108	21,6	0,16	
August	2	217	108,5	91	45,5	0,06	
September	3	71	23,67	25	8,33	0,1	
October	9	777	86,33	320	35,56	0,29	
November	10	343	34,30	251	25,10	0,33	
December	13	1.462	112,46	558	42,92	0,42	
2013	62	4.050		1.720			
January	10	2.462	246,2	1.221	122,1	0,32	2014
February	12	976	81,33	360	30	0,4	
March	7	1.295	185	553	79	0,23	
April	18	1.305	72,5	597	33,17	0,6	
May	25	2.358	94,32	1.149	45,96	0,81	
June	36	3.905	108,47	1.420	39,44	1,2	
July	34	2.126	62,53	1.043	30,68	1,1	
August	24	2.213	92,21	1.313	54,71	0,77	
September	37	4.335	117,16	1.730	46,76	1,23	
October	34	2.583	75,97	1.021	30,03	1,1	
November	33	3.995	121,06	1.682	50,97	1,1	
December	16	1.315	82,19	725	45,31	0,52	
2014	286	28.868		12.814			
Total	435	36.178		16.063			

Appendix 12: Dataset - GuardianData

GuardianData



	Tweets	Retweets	Average Retweet	Favorite	Average Favorite	Relevant Tweet/Day	Years
July	0	0	0	0	0	0,00	2011
August	5	35	7,00	31	6,2	0,20	
September	8	32	4,00	33	4,13	0,27	
October	21	130	6,19	51	2,43	0,68	
November	10	45	4,5	18	1,8	0,33	
December	13	83	6,38	49	3,77	0,42	
2011	57	325		182			
January	11	55	5	66	6	0,35	2012
February	5	18	3,6	4	0,8	0,17	
March	24	136	5,67	72	3	0,77	
April	38	317	8,34	204	5,37	1,27	
May	17	118	6,94	83	4,88	0,55	
June	14	107	7,64	39	2,79	0,47	
July	30	176	5,87	106	3,53	0,97	
August	8	47	5,88	23	2,88	0,26	
September	11	87	7,91	37	3,36	0,37	
October	30	404	13,47	151	5,03	0,97	
November	21	160	7,62	101	4,8	0,70	
December	8	107	13,38	78	9,75	0,26	
2012	217	1.732		964			
January	16	153	9,5625	84	5,25	0,52	2013
February	8	102	12,75	74	9,25	0,29	
March	11	152	13,82	58	5,27	0,35	
April	11	114	10	75	6,8	0,37	
May	5	15	3	14	2,8	0,2	
June	5	45	9,00	29	5,80	0,17	
July	7	44	6,29	19	2,71	0,23	
August	4	25	6,25	14	3,5	0,13	
September	5	77	15,40	17	3,40	0,2	
October	4	45	11,25	12	3,00	0,13	
November	6	79	13,17	43	7,17	0,20	
December	8	244	30,50	92	11,50	0,26	
2013	90	1.095		531			
January	6	123	20,5	69	11,5	0,19	2014
February	4	13	3,25	10	2,5	0,1	
March	14	404	28,86	131	9	0,45	
April	13	190	14,6	66	5,08	0,4	
May	20	370	18,50	133	6,65	0,65	
June	7	141	20,14	57	8,14	0,2	
July	7	107	15,29	48	6,86	0,2	
August	17	253	14,88	72	4,24	0,55	
September	54	2.193	40,61	519	9,61	1,80	
October	30	2.464	82,13	798	26,60	1,0	
November	31	565	18,23	190	6,13	1,0	
December	21	421	20,05	177	8,43	0,68	
2014	224	7.244		2.270			
Total	588	10.396		3.947			

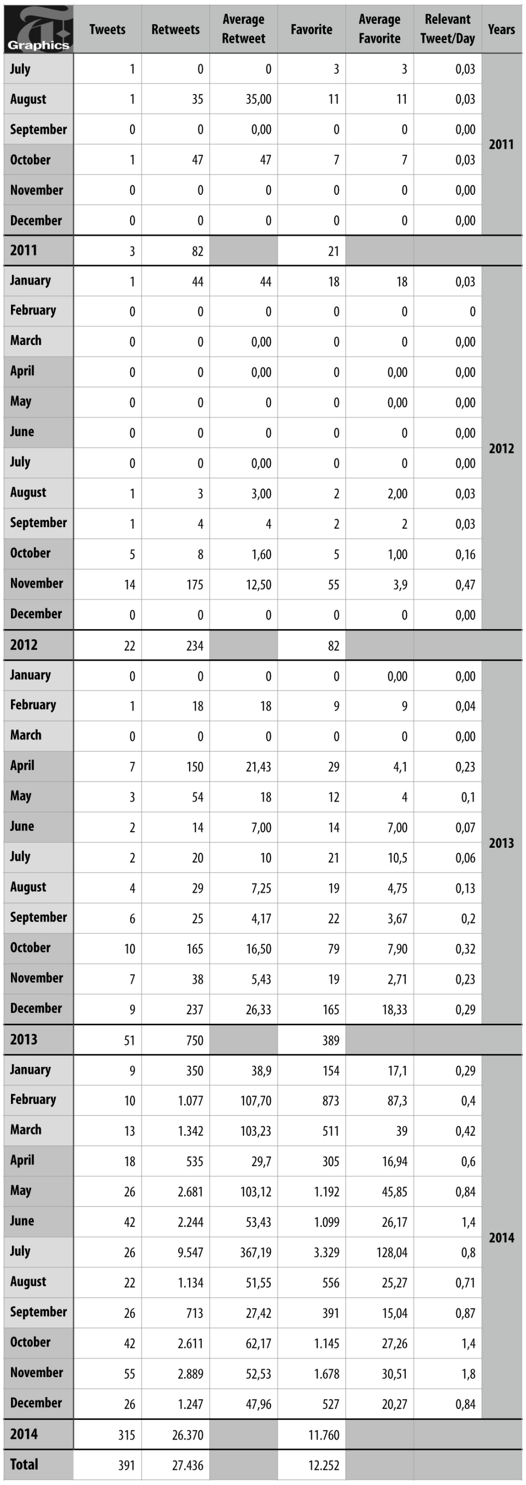
Appendix 13: Dataset - The New York Times

The New York Times

	Tweets	Retweets	Average Retweet	Favorite	Average Favorite	Relevant Tweet/Day	Years
July	2	95	47,5	22	11	0,06	**2011**
August	1	201	201	78	78	0,03	
September	0	0	0	0	0	0,00	
October	1	267	267	94	94	0,03	
November	0	0	0	0	0	0	
December	0	0	0	0	0	0	
2011	4	563		194			
January	1	125	125	65	65	0,03	**2012**
February	0	0	0	0	0	0	
March	0	0	0	0	0	0	
April	0	0	0	0	0	0	
May	1	37	37	13	13	0,03	
June	0	0	0	0	0	0	
July	0	0	0	0	0	0	
August	0	0	0	0	0	0,00	
September	8	660	82,5	229	28,625	0,27	
October	4	529	132,25	77	19,25	0,13	
November	1	69	69	41	41,0	0,03	
December	5	627	125,4	176	35,2	0,16	
2012	20	2.047		601			
January	5	620	124	316	63,2	0,16	**2013**
February	6	762	127	152	25,33	0,21	
March	2	114	57	22	11	0,06	
April	11	3.044	276,73	711	64,64	0,37	
May	9	824	91,56	237	26,3	0,3	
June	12	1.288	107,33	712	59,33	0,40	
July	7	364	52	147	21	0,23	
August	17	1.050	61,76	588	34,59	0,55	
September	17	1.009	59,35	457	26,88	0,6	
October	29	4.131	142,45	1.679	57,9	0,94	
November	26	2.237	86,04	1.227	47,19	0,87	
December	20	3.148	157,4	1.730	86,5	0,65	
2013	161	18.591		7.978			
January	23	3.193	138,83	1.942	84,43	0,74	**2014**
February	34	4.342	127,71	3.445	101,32	1,21	
March	47	9.708	206,55	5.936	126,30	1,52	
April	17	2.510	147,65	1.692	99,53	0,6	
May	77	23.941	310,92	16.586	215,40	2,48	
June	154	73.084	474,57	42.437	275,56	5,1	
July	139	50.880	366,04	27.274	196,22	4,48	
August	146	45.137	309,16	28.160	192,88	4,71	
September	149	41.239	276,77	26.245	176,14	4,97	
October	157	123.141	784,34	30.137	191,96	5,06	
November	112	31.171	278,31	19.237	171,76	3,73	
December	79	18.269	231,25	12.358	156,43	2,55	
2014	1.134	426.615		215.449			
Total	1.319	447.816		224.222			

Appendix 14: Dataset - New York Times Graphics

NYT Graphics



Graphics	Tweets	Retweets	Average Retweet	Favorite	Average Favorite	Relevant Tweet/Day	Years
July	1	0	0	3	3	0,03	2011
August	1	35	35,00	11	11	0,03	
September	0	0	0,00	0	0	0,00	
October	1	47	47	7	7	0,03	
November	0	0	0	0	0	0,00	
December	0	0	0	0	0	0,00	
2011	3	82		21			
January	1	44	44	18	18	0,03	2012
February	0	0	0	0	0	0	
March	0	0	0,00	0	0	0,00	
April	0	0	0,00	0	0,00	0,00	
May	0	0	0	0	0,00	0,00	
June	0	0	0	0	0	0,00	
July	0	0	0,00	0	0	0,00	
August	1	3	3,00	2	2,00	0,03	
September	1	4	4	2	2	0,03	
October	5	8	1,60	5	1,00	0,16	
November	14	175	12,50	55	3,9	0,47	
December	0	0	0	0	0	0,00	
2012	22	234		82			
January	0	0	0	0	0,00	0,00	2013
February	1	18	18	9	9	0,04	
March	0	0	0	0	0	0,00	
April	7	150	21,43	29	4,1	0,23	
May	3	54	18	12	4	0,1	
June	2	14	7,00	14	7,00	0,07	
July	2	20	10	21	10,5	0,06	
August	4	29	7,25	19	4,75	0,13	
September	6	25	4,17	22	3,67	0,2	
October	10	165	16,50	79	7,90	0,32	
November	7	38	5,43	19	2,71	0,23	
December	9	237	26,33	165	18,33	0,29	
2013	51	750		389			
January	9	350	38,9	154	17,1	0,29	2014
February	10	1.077	107,70	873	87,3	0,4	
March	13	1.342	103,23	511	39	0,42	
April	18	535	29,7	305	16,94	0,6	
May	26	2.681	103,12	1.192	45,85	0,84	
June	42	2.244	53,43	1.099	26,17	1,4	
July	26	9.547	367,19	3.329	128,04	0,8	
August	22	1.134	51,55	556	25,27	0,71	
September	26	713	27,42	391	15,04	0,87	
October	42	2.611	62,17	1.145	27,26	1,4	
November	55	2.889	52,53	1.678	30,51	1,8	
December	26	1.247	47,96	527	20,27	0,84	
2014	315	26.370		11.760			
Total	391	27.436		12.252			